Non-Convex Optimization for Latent Data Models : Algorithms, Analysis and Applications

Belhal Karimi

NNT : 2019SACLX040

THÈSE DE DOCTORAT

de

L'Université Paris-Saclay

École doctorale de mathématiques Hadamard (EDMH, ED 574)

Établissement d'inscription : École Polytechnique

Laboratoire d'accueil : Centre de mathématiques appliquées de Polytechnique (UMR 7641 CNRS)

Spécialité de doctorat : Mathématiques appliquées

Belhal KARIMI

Non-Convex Optimization for Latent Data Models: Algorithms, Analysis and Applications

Optimisation Non Convexe pour Modèles à Données Latentes:
Algorithmes, Analyse et Applications

Date de soutenance : 19 Septembre 2019

Lieu de soutenance : Palaiseau

Après avis des rapporteurs :

Zaid HARCHAOUI	(Pr., Department of Statistics, University of Washington)
Julien MAIRAL	(C.R., THOTH team, Inria Grenoble)

Jury de soutenance :

Aymeric DIEULEVEUT	(M.C., CMAP, École Polytechnique) Examinateur
Gersende FORT	(D.R., CNRS, IMT Toulouse) Présidente
Alexandre GRAMFORT	(Pr., PARIETAL team, Inria Saclay) Examinateur
Zaid HARCHAOUI	(Pr., Department of Statistics, UW) Rapporteur
Marc LAVIELLE	(D.R., XPOP team, Inria Saclay) Codirecteur de thèse
Julien MAIRAL	(C.R., THOTH team, Inria Grenoble) Rapporteur
Eric MOULINES	(Pr., CMAP, École Polytechnique) Codirecteur de thèse
Alessandro RUDI	(C.R., SIERRA team, Inria - DI ENS) Examinateur

A la mémoire de Stanislas Brien.

A mes amis.

"All different, all equal."

Remerciements

Je souhaiterais commencer par remercier mes directeurs de thèse pour m'avoir accompagné avec enthousiasme et loyauté tout au long de ces trois années. Marc, je te suis reconnaissant pour ta disponibilité, ta compréhenssion et ta bonne humeur au cours de cette thèse. Tu m'as enseigné l'art des statistiques et m'a plus que grandement sensibilisé aux sciences de la vie. Ton travail acharné en-dedans et en-dehors du CMAP ont énormément de sens, c'est admirable. Eric, la quantité de connaissances que tu m'as transmis ainsi que la méthodologie de recherche que tu m'as enseigné n'ont pas de prix. Ce que tu m'as apporté pendant ces trois ans sont la raison pour laquelle j'ai effectué cette thèse et je t'en remercie. Vos conseils me suivront toute ma vie.

Je remercie également mes rapporteurs Julien Mairal et Zaid Harchaoui pour avoir accepté de relire ce manuscrit ainsi qu'à l'ensemble de mon jury pour avoir accepté l'invitation.

Je remercie ensuite l'ensemble de l'équipe Xpop Inria. Hanadi, ta serviabilité et ta gentillesse ont rendu cette thèse bien plus agréable. Je remercie Julie et Erwan pour leur soutien. Julie pour sa bonne humeur et sa motivation communicatives. Erwan pour m'avoir donné l'opportunité d'enseigner en formation continue: ce n'était pas toujours évident mais j'en retiens beaucoup de choses.

I would like to thank Hoi To Wai for his guidance and his friendship along these years. I am truly honoured to have met you. You are an outstanding researcher and human being. I wish to visit you soon in HK. Je remercie Emmanuelle Comets pour nos séances de travail parisiennes. En ésperant collaborer d'avantage avec toi.

I would like to thank Dr. Dmitry Vetrov and Dr. Dmitry Kropotov for their warm welcome at the Samsung AI - HSE Lab in Moscow. My desire to study in the field of Bayesian Deep Learning was fulfilled in the best possible way thanks to you and the team. Thank you to Anton for his daily support for maths, programming and food. Irina, Nadia and Ekaterina, thank you for the joy and the fun you brought to this visit. Katia, my amazing guide through the streets of Moscow, a wonderful encounter.

Wilfried, le père, le footballer, le motard, le pote, merci pour ces quelques années. Une belle rencontre. Merci à Maud, Nassera, Alexandra L. et Alexandra N. pour leur bonne

humeur et efficacité.

Merci à Pierre pour son amitié et ses franches rigolades. Les meilleures pauses étaient celles que l'on passait ensemble même si elles auraient parfois pu me coûter mon inscription en thèse. Merci à Sylvain par la même occasion qui a fait le recrutement de sa vie sur ce coup. Je préfère te prévenir, l'élève surpassera très bientôt le maître.

Zoltan, thank you for your availability to discuss anything, even not kernel related. I wish you all the best for your career and HDR. Par ailleurs, je n'aurais sûrement pas vécu cette thèse de la même façon sans la compagnie des autres doctorants du CMAP. Luca, ta rencontre est une chance, plus qu'un collègue, un ami. Ta passion pour **la** mathématiques (faute très classique pour un Italien qui se risque au français) m'a souvent fait rire, mais elle a certainement déteint sur moi pour le meilleur. Cedric, Massil, Romain et Gustaw, sans vous je n'aurai jamais su que j'étais nul au hashcode, c'est bon à savoir. Cedric merci pour ce diner brésilien auquel tu m'as invité mais auquel je n'ai pas pu assister (tu comprendras). Geneviève, une inspiration constante sur tous les plans. Mathilde, ma confidente, je te remercie pour nos discussions existentielles et philosophiques. Nos bavardages autour d'un bol de chicoré m'ont aidé à me construire.

Merci à Rémi, Frederic, Aude, Corentin H. et Corentin C. pour avoir été un si bon public. Si jamais mon *one-man* show voit le jour, vous en percevrez des royalties.

Une pensée particulière pour Jaouad qui m'a bien rassuré quant à la quantité de travail et à ma présence au labo. Blague à part, une belle rencontre et un chercheur hors pair.

Pensées particulières pour les relectures de Nicolas et Frederic.

Un énorme remerciement à l'équipe de basket-ball de l'X coachée par les adjudants-chefs Grégory Puma et Thierry Plat. Merci Greg pour ta rigueur et tes conseils afin d'être tout simplement la meilleure version de moi-même. Ces conseils ont été une force au labo comme sur le terrain. Merci Thierry pour ta pédagogie et le savoir basket que tu m'as transmis. Merci à toute l'équipe: Rémi, David, Marco, Anthony, Amaury, Maxime, Kader, Badr, Konstantinos, Edison et Antonio. Cette dernière année fut pleine de réussite. Le quart de championnat de France n'est qu'une premiere étape pour la saison prochaine. Je compte sur vous.

Je remercie l'équipe Data et Assurances de OuiCar pour leur accueil durant ma troisème année. Je suis particulièrement reconnaissant envers Christophe pour m'avoir donné une chance de développer mon idée dans leur locaux (et avec leurs ressources si précieuses).

Je remercie évidemment l'ensemble de mes amis. Des Ulis à Paris en passant par Orsay, vous avez été très compréhenssifs. Mon équilibre de vie et ma réussite pendant cette thèse, je les dois en grande partie à vous.

Je termine par le plus important : la famille. Merci à mes parents pour avoir toujours été

présents et pour m'avoir soutenu. Merci également à mes grands frères pour leur soutien inconditionnel. Pour finir, merci à mes tantes Mina et Dina: du pur bonheur à chaque fois que l'on se voyait.

Abstract

Many problems in machine learning pertain to tackling the minimization of a possibly non-convex and non-smooth function defined on a Euclidean space. Examples include topic models, neural networks or sparse logistic regression. Optimization methods, used to solve those problems, have been widely studied in the literature for convex objective functions and are extensively used in practice. However, recent breakthroughs in statistical modeling, such as deep learning, coupled with an explosion of data samples, require improvements of non-convex optimization procedures for large datasets. This thesis is an attempt to address those two challenges by developing algorithms with cheaper updates, ideally independent of the number of samples, and improving the theoretical understanding of non-convex optimization that remains rather limited.

In this manuscript, we are interested in the minimization of such objective functions for latent data models, *i.e.,* when the data is partially observed which includes the conventional sense of missing data but is much broader than that. In the first part, we consider the minimization of a (possibly) non-convex and non-smooth objective function using *incremental* and *online* updates. To that end, we propose and analyze several algorithms, that exploit the latent structure to efficiently optimize the objective function, and illustrate our findings with numerous applications. In the second part, we focus on the maximization of non-convex likelihood using the EM algorithm and its stochastic variants. We analyze several faster and cheaper algorithms and propose two new variants aiming at speeding the convergence of the estimated parameters.

In the first main contribution, we provide a unified framework of analysis for optimizing non-convex finite-sum problems which encompasses logistic regression and variational inference. This framework is an extension of an incremental surrogate optimization method based on the Majorization-Minimization principle and aims at minimizing an easier upper bound of the objective function at each iteration of the algorithm in an incremental fashion. Our proposed framework is proved to converge almost surely to a stationary point and in $\mathcal{O}(n/\epsilon)$ iterations to an ϵ-stationary point.

In the second main contribution, we analyze a stochastic approximation scheme where the stochastic drift term is non necessarily a gradient and with a potentially biased mean

field under two cases: the vector of random variables is either i.i.d. or a *state-dependent* Markov chain. For both cases, we provide tight non-asymptotic upper bounds, of order $\mathcal{O}(c_0 + \log(n)/\sqrt{n})$, where c_o is the potential bias of the drift term, and illustrate our findings by analyzing popular statistical learning algorithms such as the online expectation maximization (EM) algorithm and the average cost policy-gradient method.

The third main contribution deals with the maximum likelihood (ML) estimation problem. We propose and analyze fast incremental variants of the EM algorithm, as one of the most popular algorithm for inference in latent data models. We show that the incremental version of the EM is a special instance of an incremental surrogate optimization framework, and takes $\mathcal{O}(n/\epsilon)$ iterations to find an ϵ-stationary point to the ML estimation problem. We propose a faster incremental variant that takes $\mathcal{O}(n^{2/3}/\epsilon)$ iterations to find to an ϵ-stationary point and show that a recently proposed variance reduced stochastic EM method has the same iteration complexity.

The fourth main contribution of the manuscript develops a fast variant of the Stochastic Approximation of the EM algorithm to tackle the ML estimation problem in nonlinear mixed effect models. In this context, the latent structure corresponds to the random effects that are random variables associated with each sample (individual) from a population. Our proposed algorithm improves the sampling procedure, used to simulate the aforementioned individual random effects, and its performances are studied experimentally through several pharmacology applications.

The fifth main contribution deals with an incremental variant of the Stochastic Approximation of the EM algorithm whose convergence guarantees are studied both theoretically and experimentally.

The sixth main contribution presents an R package, extending a current version of the *saemix* R package, useful for training noncontinuous data models, such as categorical or time-to-event, using the SAEM algorithm. We illustrate the convenience of our extended package on two simple numerical examples and provide (and explain) the lines of code to perform maximum likelihood estimation.

Keywords: stochastic approximation, non-convex optimization, finite-sum, large-scale, latent data, EM, MCMC, incremental, online.

Résumé

De nombreux problèmes en Apprentissage Statistique consistent à minimiser une fonction non convexe et non lisse définie sur un espace euclidien. Par exemple, les problèmes de maximisation de la vraisemblance et la minimisation du risque empirique en font partie. Les algorithmes d'optimisation utilisés pour résoudre ce genre de problèmes ont été largement étudiés pour des fonctions convexes et grandement utilisés en pratique. Cependant, l'accrudescence du nombre d'observations dans l'évaluation de ce risque empirique ajoutée à l'utilisation de fonctions de perte de plus en plus sophistiquées représentent des obstacles. Ces obstacles requièrent d'améliorer les algorithmes existants avec des mis à jour moins coûteuses, idéalement indépendantes du nombre d'observations, et d'en garantir le comportement théorique sous des hypothèses moins restrictives, telles que la non convexité de la fonction à optimiser.

Dans ce manuscrit de thèse, nous nous intéressons à la minimisation de fonctions objectives pour des modèles à données latentes, *i.e.,* lorsque les données sont partiellement observées ce qui inclut le sens conventionnel des données manquantes mais est un terme plus général que cela. Dans une première partie, nous considérons la minimisation d'une fonction (possiblement) non convexe et non lisse en utilisant des mises à jour *incrémentales* et *en ligne*. Nous proposons et analysons plusieurs algorithmes à travers quelques applications. Dans une seconde partie, nous nous concentrons sur le problème de maximisation de vraisemblance non convexe en ayant recourt à l'algorithme EM et ses variantes stochastiques. Nous en analysons plusieurs versions rapides et moins coûteuses et nous proposons deux nouveaux algorithmes du type EM dans le but d'accélérer la convergence des paramètres estimés.

La première contribution de cette thèse est un cadre unifié d'analyse pour l'optimisation d'une grande somme finie de fonction non convexes qui inclut la régression logistique et l'inférence variationnelle. Ce cadre est une extension d'une méthode d'optimisation par fonction surrogate incrémentale basée sur le principe de Majorisation-Minimisation et vise à minimiser, à chaque itération, une fonction majorante plus simple à optimiser de manière incrémentale. Nous prouvons que notre méthode converge presque sûrement vers un point stationnaire et avec une complexité de $\mathcal{O}(n/\epsilon)$ itérations vers un point ϵ-stationnaire.

Dans la deuxième contribution, nous analysons un schéma d'approximation stochastique dont le terme de dérive n'est pas nécessairement un gradient et dont le champ moyen peut présenter un biais. Deux cas sont distingués: le vecteur de variables aléatoires qui définit le terme de dérive est i.i.d. ou une chaîne de Markov. Dans les deux cas, nous déterminons des limites supérieures étroites, d'ordre $\mathcal{O}(c_0 + \log(n)/\sqrt{n})$, avec c_o le biais potentiel du terme de dérive, et illustrons nos conclusions à travers l'analyse de l'algorithme EM en ligne et l'algorithme de descente de gradient sur les politiques en apprentissage par renforcement.

La troisième contribution aborde le problème de maximisation de vraisemblance. Nous y proposons et analysons des variantes incrémentales et rapides de l'EM, considéré comme l'un des algorithmes les plus populaires pour de l'inférence au sein de modèles à données latentes. Nous montrons que l'EM incrémentale est une instance d'une cadre d'optimisation par fonction surrogate incrémentale and requiert $\mathcal{O}(n/\epsilon)$ itérations pour trouver un point ϵ-stationnaire au problème de maximum de vraisemblance. Nous proposons également une version rapide de ce dernier algorithme qui requiert $\mathcal{O}(n^{2/3}/\epsilon)$ itérations pour trouver un point ϵ-stationnaire.

La quatrième contribution de cette thèse développe une version rapide de l'algorithme SAEM (Stochastic Approximation of the EM) afin d'estimer les paramètres de population dans des modèles non linéaires à effets mixtes. Ici, la structure latente correspond aux effets aléatoires qui sont des variables aléatoires associées à chaque échantillon (individu) d'une même population. Notre algorithme améliore la procédure d'échantillonage des effets aléatoires et ses performances sont étudiées de manière expérimentales à travers plusieurs applications en pharmacologie.

La cinquième contribution développe une version incrémentale de l'algorithme SAEM dont les propriétés de convergence asymptotique sont étudiées sur le plan théorique et pratique.

La sixième contribution présente l'utilisation d'un package R, construit autour d'un package existant nommé *saemix*, utile à l'entrainement de modèles non continus, tels que les modèles catégoriques ou de survie, utilisant l'algorithme SAEM. Nous illustrons son utilisation sur deux examples numériques simples et fournissons (et expliquons) les lignes de code à executer.

Mots clés: approximation stochastique, optimisation non convexe, somme-finie, grande-echelle, données latentes, EM, MCMC, incremental, en ligne.

Contents

Nomenclature

$\mathbb{R}$	Set of real numbers	Ensemble des réels
X	Non-empty set	Ensemble non vide
$\mathcal{X}$	σ-algebra on X	Tribu (ou σ-algèbre)
$(\mathsf{X}, \mathcal{X})$	Measurable space	Espace mesurable
y	Observations	Observations
z	Latent variables	Variable latentes
ψ	Individual parameters	Paramètres individuels
η	Random effects	Effets aléatoires
n	Number of observations	Nombre d'observations
$\boldsymbol{\theta}$	Parameters	Paramètres
Θ	Parameters Set	Ensemble de Paramètres
$[\![1, n]\!]$	Set $\{1, \dots, n\}$	Ensemble $\{1, \dots, n\}$
$\|\cdot\|$	Euclidean norm	Norme Euclidienne
$\langle \cdot \,\vert\, \cdot \rangle$	Inner product in the Euclidean space	Produit scalaire dans l'expace Euclidien
$g(y, \boldsymbol{\theta})$	Incomplete likelihood	Vraisemblance incomplète
$f(z, y, \boldsymbol{\theta})$	Complete likelihood	Vraisemblance complète
$p(z\|y, \boldsymbol{\theta})$	Posterior distribution	Distribution a Posteriori
$\mu(\cdot)$	σ-finite measure	Measure σ-finie
$\mathbb{E}[\cdot]$	Expectation	Espérance
$S(\cdot)$	Sufficient statistics	Statistiques suffisantes
$\mathcal{L}(\cdot)$	Objective function	Fonction Objective
$\mathrm{R}(\cdot)$	Regularizer	Fonction de régularisation
$\nabla\mathcal{L}(\boldsymbol{\theta})$	Gradient of $\mathcal{L}$ at $\boldsymbol{\theta}$	Gradient de $\mathcal{L}$ en $\boldsymbol{\theta}$
$\mathcal{L}'(\boldsymbol{\theta}, \boldsymbol{d})$	Directional derivation of $\mathcal{L}$ along $\boldsymbol{d}$	Dérivée directionnelle de $\mathcal{L}$ selon $\boldsymbol{d}$
$\mathrm{J}^{\mathcal{L}}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$	Jacobian of $\mathcal{L}$ at $\boldsymbol{\theta}$	Jacobienne de $\mathcal{L}$ en $\boldsymbol{\theta}$
$\mathrm{H}^{\mathcal{L}}_{\boldsymbol{\theta}}(\boldsymbol{\theta})$	Hessian of $\mathcal{L}$ at $\boldsymbol{\theta}$	Hessienne de $\mathcal{L}$ en $\boldsymbol{\theta}$

Contributions and thesis outline

This thesis is divided into two parts. chapters 3 and 4 discuss optimization of non-convex objective function, while Chapters 5 to 7 concern maximum likelihood estimation methods and their applications to pharmacology. Chapter 8 details a tutorial of the algorithm presented in Chapter 6, developed using R programming language. Each chapter can be read independently of the others.

Chapter 1: In the opening chapter, we introduce the primary optimization problem of our interest and give a short introduction to non-convex optimization, statistical learning and stochastic approximation which are the main topics of this manuscript. We overview *state-of-the-art* results found in the literature and emphasize on the statistical analysis gap that this manuscript is attempting to bridge. We also introduce our main field of applications that is pharmacology and its framework of analysis,*i.e.,* Mixed Effects Models.

Chapter 3: This chapter considers a minimization by incremental stochastic surrogate method to optimize a finite-sum objective function. It extends the work of Mairal [2015a] by deriving *Monte-Carlo* approximations of the surrogate functions minimized at each iteration. Both finite-time and asymptotic analyses are provided and illustrated through several numerical applications.

Chapter 4: This chapter develops the first analysis results regarding a stochastic approximation scheme, with potentially biased mean-field, used to find the roots, or the extrema, of a non-convex function. The novelty of the analysis, in the biased case, rests upon a new technique based on the Poisson equation. Tight upper bounds are provided in this chapter and thorough analyses of the online Expectation-Maximization (EM) and the policy gradient algorithms, as special instances of this scheme, are displayed.

Chapter 5: This chapter provides non-asymptotic convergence rates for several incremental variants of the EM algorithm. We offer two complementary views for the global convergence of incremental EM methods – one focuses on the parameter space, and the other on the sufficient statistics space. On one hand, the EM method can be studied as a *majorization-minimization* (MM) method in the parameter space. On the other hand, the EM method can be studied as a *scaled-gradient method* in the sufficient statistics space. Several numerical applications illustrate our findings.

Chapter 6: This chapter introduces inference in nonlinear mixed effects models using the Stochastic Approximation of the EM algorithm and propose a fast variant of the latter using a faster Markov Chain Monte Carlo (MCMC) sampler. The main contributions in this chapter are the construction of independent proposals, for both continuous and noncontinuous data models, used in an MCMC procedure. Numerical applications on a pharmacokinetics model and a time-to-event example confirm the advantage of our method.

Chapter 7: This chapter introduces an incremental variant of the Stochastic Approximation of the EM (SAEM) and studies its asymptotic guarantees. Findings are illustrated through several pharmacokinetics-pharmacodynamics examples.

Chapter 8: In this chapter, a tutorial for using the SAEM algorithm (presented in Chapters 6 and 7) is developed in the R programming language. We extend an existing R package for noncontinuous data models, such as categorical or time-to-event data models, and provide the correct syntax to execute the method.

Chapter 9: This chapter concludes the thesis by summarizing our contributions and describing possible extensions.

Papers related to this manuscript are listed bellow:

- **Chapter 3** is based on *A Doubly Stochastic Surrogate Optimization Scheme for Non-convex Finite-sum Problems*, B. Karimi, H.T. Wai and E. Moulines, 2019 [Karimi et al., 2019b].
- **Chapter 4** is based on *Non-asymptotic Analysis of Biased Stochastic Approximation Scheme* , B. Karimi, B. Miasojedow, E. Moulines and H.T. Wai, Conference on Learning Theory, COLT 2019 [Karimi et al., 2019a].
- **Chapter 5** is based on *On the Global Convergence of (Fast) Incremental variants of the EM* , B. Karimi, H.T. Wai, M. Lavielle and E. Moulines, Advances in Neural Information Processing Systems, NeurIPS 2019 [Karimi et al., 2019c].
- **Chapter 6** is based on *f-SAEM: A fast Stochastic Approximation of the EM algorithm* , B. Karimi, M. Lavielle and E. Moulines, Computational Statistics and Data Analysis, CSDA 2018 [Karimi et al., 2020] and *Efficient Metropolis-Hastings sampling for nonlinear mixed effects models* , B. Karimi and M. Lavielle, Springer Series Statistics and Data Science: new challenges, new generations, BAYSM 2018 [Karimi and Lavielle, 2018].

Chapter 1

Introduction

Abstract: *This introductory chapter describes the objectives of this thesis and introduces the key areas that will be studied in the following chapters. We give here a thorough overview of the literature related to those fields and emphasize on the gap that this manuscript is attempting to bridge. Several important assumptions and definitions, made throughout this document, are presented and motivated in this chapter in order to become familiar with non-convex optimization, stochastic approximation and latent data models. The closing section develops a specific instance of latent data models called mixed effects models and its application to pharmacology, as the main field of application of our team, Xpop, at Inria. This brief overview is intended to convey the flavor of the work contained herein as we will provide additional backgrounds and motivations later in each chapter.*

Contents

1.1 Statistical Learning

The field of mathematical modeling has been central to human endeavor in order to have a better understanding of the world, with applications ranging from physics to social sciences. In particular, for handling large datasets and model complex phenomena, statistical learning is considered as one of its most important subfield of our modern time. It can be viewed as a principled approach for extracting useful information from data that can be exploited to carry out tasks such as prediction. Generally, it consists of a *modeling* phase, where a model function is designed in a given model search space — in this thesis, we restrict ourselves to parametric models where the search space is the set of parameters — and a *training* or *optimization* phase where, given input-output observation pairs, the model is fitted to describe the data as well as possible. We now give a rigorous formulation of the ideas introduced above.

Mathematical formulation Consider the input-output pair of random variables (X, Y) taking values in arbitrary input set $\mathsf{X} \subset \mathbb{R}^p$ and arbitrary output set $\mathsf{Y} \subset \mathbb{R}^q$. For instance, X is a matrix of covariates describing a hospital patient (age, weight, etc.) and Y describes his or her hepatitis C viral load. We denote by $\mathcal{P}$, the distribution from which this input-output pair is drawn. As mentioned above, the *modeling* phase consists of finding a measurable function $M_{\boldsymbol{\theta}} : \mathsf{X} \mapsto \mathsf{Y}$ that is in our case a parametric function of parameter $\boldsymbol{\theta} \in \mathbb{R}^d$. This function is commonly called the *predictor* and its performance is measured using a *loss* function $\ell : \mathsf{Y} \times \mathsf{Y} \mapsto \mathbb{R}$ where $\ell(y, y')$ is the loss incurred when the true output is y whereas y' is predicted. Then, the *training* phase boils down to computing the following quantity:

$$\arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \overline{\mathcal{L}}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \left\{ \mathcal{L}(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta}) \right\} \quad \text{with} \quad \mathcal{L}(\boldsymbol{\theta}) = \mathbb{E}_{(x,y)\sim\mathcal{P}}\left[\ell(y, M_{\boldsymbol{\theta}}(x)) \right] , \quad (1.1.1)$$

where ℓ is a possibly non-convex loss function depending on some observed data, $\mathcal{L}$ is the so-called *population risk* and $\mathrm{R}(\cdot)$ is a penalization term that imposes structure to the solution and is possibly non-smooth.

Throughout this thesis, we are interested in models where the input-output relationship is not completely characterized by the observed $(x, y) \in \mathsf{X} \times \mathsf{Y}$ pairs in the training set alone, but also depends on a set of unobserved latent variables $z \in \mathsf{Z} \subset \mathbb{R}^m$. Those models are called Latent Data Models and are formally introduced in Section 1.3. They include the incomplete data framework, *i.e.,* some observations are missing, but are far broader than that: for example, the latent structure could stem from the unknown labels in mixture models or hidden states in Hidden Markov Models. In all those cases, a simulation step is required to complete the observed data with realizations of the latent variables. The latter simulation step plays a key role in this manuscript and is thoroughly addressed in

each chapter. Formally, this specificity in our setting implies extending the loss function ℓ to accept a third argument as follows:

$$\ell(y, M_{\boldsymbol{\theta}}(x)) = \int_{\mathsf{Z}} \ell(z, y, M_{\boldsymbol{\theta}}(x)) \mathrm{d}z \ . \tag{1.1.2}$$

Note that, for the sake of notation simplicity, we use the same name for both loss functions defined on different spaces. Finally, we consider examples where the function $\mathcal{L}$ is smooth in the following sense:

Definition 1.1 *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth if and only if it is differentiable and its gradient is L-Lipschitz-continuous,* i.e., *for all $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$:*

$$\|\nabla f(\theta) - \nabla f(\vartheta)\| \leq L \|\theta - \vartheta\| \ . \tag{1.1.3}$$

Traditionally, most of the focus in statistical learning has been on developing convex loss functions ℓ and algorithms such as SVM or exponential-family graphical models. However, many important problems, such as computer vision and natural language processing, cannot be formulated as convex optimization or will be more computationally expensive than their non-convex counterparts. Indeed, although convexity can be seen as a virtue, it can also be regarded as a limitation in the complexity of the model trained to solve a given problem. For instance, the latent variable models, mentioned above as a large family of probabilistic graphical models, require non-convex optimization and are useful to tackle tasks such as identification in Gaussian mixture models (useful in many domain of applications), that could not be dealt with a convex model.

The increase in dimension/sample size and the complexity of the tasks force the statistical community to develop simpler algorithms, with a complexity at most $\mathcal{O}(n)$ where n is either the dimension or the number of observations, yet fit more sophisticated and highly non-convex models. This matter is extensively addressed in [Bottou and Bousquet, 2008] and is at the origin of the expansion of the non-convex optimization field.

1.2 Non-convex Optimization

Non-convex optimization problems arise frequently in machine learning, including feature selection, structured matrix learning, mixture modeling, and neural network training. In all those cases, the function $\mathcal{L}$ defined in the optimization objective (2.1.1) is non-convex. Convexity of a function f is defined as follows:

Definition 1.2 *A function $f : \mathbb{R}^d \mapsto \mathbb{R}$ is said to be convex if for all $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$ and all $\lambda \in (0, 1)$:*

$$f((1-\lambda)\theta + \lambda\vartheta) \leq (1-\lambda) f(\theta) + \lambda f(\vartheta) \ . \tag{1.2.1}$$

Moreover, it is said to be μ-strongly convex if for all $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$ and $\mu > 0$:

$$f(\theta) \geq f(\vartheta) + \nabla f(\vartheta)^T(\theta - \vartheta) + \frac{\mu}{2}\|\theta - \vartheta\|^2 \tag{1.2.2}$$

In this manuscript, we are interested in the constrained formulation of this optimization problem. Thus, the parameters vector $\boldsymbol{\theta}$ belongs to a convex set $\Theta \subset \mathbb{R}^d$ in the following sense:

Definition 1.3 *A set Θ is said to be convex if for all $(\theta, \vartheta) \in \Theta^2$ and all $\lambda \in (0, 1)$:*

$$(1 - \lambda)\theta + \lambda\vartheta \in \Theta \; . \tag{1.2.3}$$

Differentiability of the objective function on a constrained set is handled by introducing the following concept of directional differentiability (which includes the differentiability notion):

Definition 1.4 *For any function $f : \Theta \to \mathbb{R}$, $f'(\boldsymbol{\theta}, \boldsymbol{d})$ is the directional derivative of f at $\boldsymbol{\theta}$ along the direction $\boldsymbol{d}$,* i.e.,

$$f'(\boldsymbol{\theta}, \boldsymbol{d}) := \lim_{t \to 0^+} \frac{f(\boldsymbol{\theta} + t\boldsymbol{d}) - f(\boldsymbol{\theta})}{t} \; . \tag{1.2.4}$$

Analyzing the convergence of an optimization algorithm which is said to be *convex* (resp. *non-convex*) if the objective (2.1.1) is *convex* (resp. *non-convex*), usually implies a sub-optimality condition as the convergence criterion of interest. For instance, for convex functions, we use $|\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)|$ (or $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$) as such condition. We denote by $\boldsymbol{\theta}^*$ the optimal solution that can efficiently be found in the convex case. Consequently, when finding such optimal solution is hard, as in the non-convex case, such convergence criterion can not hold. We then use the quantity $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2$, as advocated in [Nesterov, 2004] and [Ghadimi and Lan, 2013], to evaluate the stationarity of the algorithm iterates. Thus, the following definition is important throughout our analysis:

Definition 1.5 *A point $\boldsymbol{\theta}^*$ is said to be ε-stationary if $\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|^2 \leq \varepsilon$. A stochastic iterative algorithm is said to achieve ε-stationarity in $\Gamma > 0$ iterations if $\mathbb{E}[\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(R)})\right\|^2] \leq \varepsilon$, where the expectation is over the stochasticity of the algorithm.*

We give two formulations, found in the literature, of such results in the convex and non-convex case to give a sense of the kind of bounds one can obtain to characterize stationarity of the algorithm iterates. Consider the simple following unconstrained and un-regularized optimization problem that consists of finding the parameter $\boldsymbol{\theta}^* \in \mathbb{R}^d$ such that:

$$\boldsymbol{\theta}^* := \arg\min_{\boldsymbol{\theta} \in \mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \; , \tag{1.2.5}$$

where $\mathcal{L} : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth. In the convex case, as mentioned above, a common suboptimality condition is $\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^*)$. For the Gradient Descent algorithm, an upper bound on this quantity is expressed as follows:

Proposition 1 *(Convergence of gradient descent [Nesterov, 2004] for convex functions). Consider the simple gradient descent scheme, with constant stepsize, that starts from an initial $\boldsymbol{\theta}^{(0)}$ and compute the sequence of iterates $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ as follows:*

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}) \,, \tag{1.2.6}$$

where $\mathcal{L}$ is a convex and L-smooth function on $\mathbb{R}^d$. Let the stepsize $\gamma = 1/L$, then the sequence of iterates $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ satisfies:

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^*) \leq \frac{L \left\| \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(0)} \right\|^2}{k+1} \,. \tag{1.2.7}$$

Moreover, if $\mathcal{L}$ is μ- strongly convex we have

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^*) \leq (1 - \mu/L)^{k+1} [\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^*)] \tag{1.2.8}$$

In the non-convex case, Ghadimi and Lan [2013] derived the first finite-time analysis of the well known SGD algorithm. A typical analysis trick for non-convex problems is to adopt a stopping rule. Consider the random variable Γ, playing the role of a termination point, distributed according to a given probability mass $P_\Gamma(\cdot)$, then the finite-time analysis is done at iteration Γ. See an example of such result:

Proposition 2 *(Convergence of stochastic gradient descent [Ghadimi and Lan, 2013] for non-convex functions). Consider the initial value $\boldsymbol{\theta}^{(0)}$, a termination point Γ drawn from a probability mass function $P_\Gamma(\cdot)$ supported on $\{1, \cdots, K\}$ with K an iteration limit and the following updates for $k \in [\![1, \Gamma]\!]$:*

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma_{k+1} H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) \,, \tag{1.2.9}$$

where $\{X_{k+1}\}_{k<\Gamma}$ are i.i.d., zero-mean random vectors and $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ is a noisy unbiased estimate of the gradient $\nabla \mathcal{L}(\boldsymbol{\theta}^{(k)})$ and $\mathcal{L}$ is L-smooth and a (possibly) non-convex function on $\mathbb{R}^d$. Assume that

$$\mathbb{E}[H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})] = \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}) \quad \textit{and} \quad \mathbb{E}[\left\| H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) - \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}) \right\|^2] \leq \sigma^2 \tag{1.2.10}$$

and that the stepsize $\gamma_k < 1/2L$, then, for any $N > 1$, the sequence of iterates

$\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ satisfies:

$$\frac{1}{L}\mathbb{E}[\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(\Gamma)})\right\|^2] \leq \frac{D_{\mathcal{L}}^2 + \sigma^2\sum_{k=1}^N \gamma_k^2}{\sum_{k=1}^N (2\gamma_k - L\gamma_k^2)} \tag{1.2.11}$$

with $D_{\mathcal{L}} = \sqrt{2(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^*)/L}$. The expectation is taken over the stochasticity of the algorithm.

In particular, we extend in Chapter 4 the result above for a drift term that is not necessarily a gradient and possibly a biased estimator of the mean field.

1.2.1 Empirical Risk Minimization

In general, as the data generating distribution $\mathcal{P}$ is often unknown, n pair $((y_i, x_i), i \in [\![1, n]\!])$ of independent observations, also called *training examples*, are considered in the optimization procedure (2.1.1). Based on the empirical risk minimization (ERM) principle [Vapnik, 2013], the optimization problems involve a data fitting loss function $\mathcal{L}$, also known as the *empirical risk*, averaged over those sample points. Namely, the objective function, without penalization, reads:

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^n \ell(y_i, M_{\boldsymbol{\theta}}(x_i)) \\ &= n^{-1}\sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}) \ .\end{aligned} \tag{1.2.12}$$

where n is the number of observations, and $\mathcal{L}_i$ is the loss associated to the i-th observation. The regularized ERM variant consists of adding a possibly non-smooth penalty $\mathrm{R}(\boldsymbol{\theta})$ as introduced in (2.1.1). For instance for some observation $y \in \mathsf{Y}$ and some prediction $y' \in \mathsf{Y}$ usual losses are the quadratic loss $\ell(y, y') = \|y - y'\|^2/2$ for regression task and loss of the form $\ell(y, y') = \mathbb{1}_{\{yy'<0\}}$ for a binary classification task where we recall that the prediction y' depends on observed covariates x, a model $M_{\boldsymbol{\theta}}(\cdot)$ and possibly some latent variables z.

In the convex case, many well-known deterministic methods such as Gradient Descent (GD), accelerated gradient methods and Newton's methods are used to perform the optimization task, see [Bertsekas, 1999, Boyd and Vandenberghe, 2004, Nesterov, 2004] and the references therein. However, each of these methods are computationally involved as they require a full pass over the dataset at each iteration. To deal with a large number n of training points, stochastic and incremental first-order and second-order optimization methods have been popular and widely studied when the objective is convex, see [Defazio et al., 2014, Mairal, 2015b, Roux et al., 2012, Vanli et al., 2018], as cheaper per-iteration cost algorithms, at the price of a higher memory cost. For non-convex objective functions, deterministic [Agarwal et al., 2017, Carmon et al., 2017] and stochastic [Allen-Zhu

and Hazan, 2016, J. Reddi et al., 2016] methods have also been developed to reach an ε-stationary point. It is important to note that in the non-convex case, an ε-stationary point could be a saddle point. Many important works are being done in the direction of escaping those saddle points to ensure reaching a local minima of (2.1.1), as in [Reddi et al., 2018, Royer and Wright, 2018, Xu et al., 2018], but they are outside the scope of this thesis.

A popular class of algorithms for solving the minimization of a non-convex composite function are majorization-minimization [Lange, 2016] techniques which iteratively approximate the composite nonconvex function by a majorizing function that is easy to minimize. For instance, most techniques, such as GD, use a quadratic convex majorizer that can be optimized efficiently. An illustration of that concept is provided Figure 2.1.

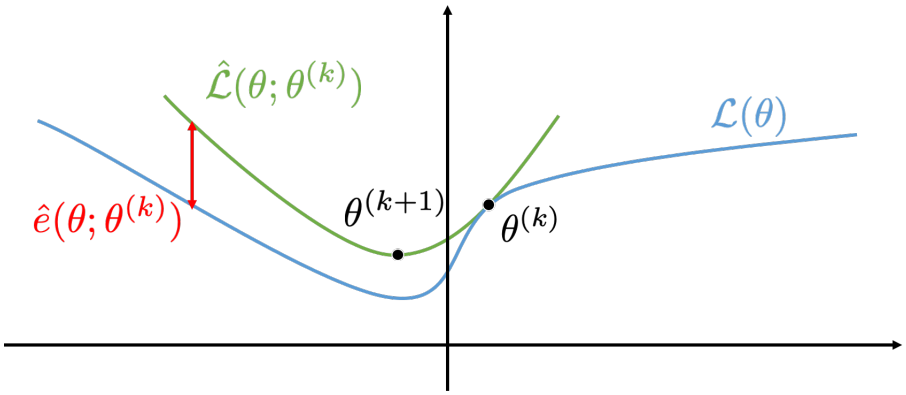

Figure 1.1 – Majorization-Minimization principle.

Note that, at iteration k, the objective $\mathcal{L}(\boldsymbol{\theta})$ is upper bounded by a tight surrogate $\hat{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ at the current estimate $\boldsymbol{\theta}^{(k)}$. An incremental variant exploiting the finite sum structure of the problem has been developed by [Mairal, 2015b] and is extended Chapter 3. In particular, our extension builds upon a doubly stochastic scheme: the first level of stochasticity stems from the choice of the individual objective component $\mathcal{L}_i$ for $i \in [\![1, n]\!]$, thus exploiting the finite-sum structure of the problem, while the second level of stochasticity profits from the latent structure of the problem to build suitable surrogate functions. A well-known example of this incremental framework is the incremental Expectation-Maximization (EM) algorithm developed in the pioneering paper by Neal and Hinton [1998]. The authors take advantage of the latent structure of the problem, introduced in the opening section, to build *easy-to-optimize* majorizing surrogates (see Chapter 5 for a thorough presentation of this approach). In this thesis, we will focus on algorithms that use incremental first-order oracles.

Definition 1.6 *(Incremental first-order oracle) For a given function $f : \Theta \mapsto \mathbb{R}$ with a finite-sum structure, an increment first-order oracle takes an index $i \in [\![1, n]\!]$ and a parameter estimate $\boldsymbol{\theta} \in \Theta$ and returns the values of $f_i(\boldsymbol{\theta})$ and/or its gradient $\nabla f_i(\boldsymbol{\theta})$.*

In the current state of the literature, these algorithms are favored as they require only a small amount of first-order information at each iteration.

Prior Work In the (possibly strongly) convex case, Stochastic Gradient Descent (SGD) has been at the center of huge progress in the past decade. Many incremental variants [Bertsekas, 2011] have been developed since its introduction in the seminal work [Robbins and Monro, 1951]. Among them, a class of variance reduced algorithms are proven to achieve faster rates for convex objective. For instance [Defazio et al., 2014, Roux et al., 2012] have developed fast incremental algorithms that achieve linear convergence rates for strongly convex functions. The Stochastic Variance Reduced Gradient (SVRG) [Johnson and Zhang, 2013] is another variance reduced method which displays a lower storage requirement compared to the latter methods. Moreover, a study of lower bounds for composite function optimization problem has been done in [Agarwal and Bottou, 2014], yet the literature remains rather poor for the non-convex setting.

In the non-convex case, several important works [Bottou, 1991, Kushner and Clark, 2012] develop asymptotic convergence of incremental variants of SGD to a stationary point. The first non asymptotic convergence rate of SGD in [Ghadimi and Lan, 2013] ensures an ε-stationary point in $\mathcal{O}(1/\varepsilon^2)$ iterations, see Table 2.1. Incremental variants are also analyzed in [Ghadimi et al., 2016] and in particular the SVRG is known to achieve an ε-stationary point in $\mathcal{O}(n^{2/3}/\varepsilon)$ in [Reddi et al., 2016a]. Those results are relevant in the sense that they separate themselves from a local convexity assumption and tackle the global convergence behavior of optimization methods in the non-convex setting. This manuscript follows the same spirit.

Algorithm	Gradient	Non-gradient	MC	Step.
SGD	$\mathcal{O}(1/\varepsilon^2)$ [Ghadimi and Lan, 2013]	TBD	×	γ_k
GD	$\mathcal{O}(n/\varepsilon)$ [Nesterov, 2004]	TBD	×	γ
SVRG/SAGA	$\mathcal{O}(n^{2/3}/\varepsilon)$ [Reddi et al., 2016b]	$\mathcal{O}(n^{2/3}/\varepsilon)$ Chap. 5	×	γ_k
MISO	$\mathcal{O}(n/\varepsilon)$ Chap. 3	$\mathcal{O}(n/\varepsilon)$ Chap. 3	×	—
MISSO	$\mathcal{O}(n/\varepsilon)$ Chap. 3	$\mathcal{O}(n/\varepsilon)$ Chap. 3	✓	—
Biased SA	$\mathcal{O}(c_0 + \frac{\log(n)}{\varepsilon\sqrt{n}})$ Chap. 4	$\mathcal{O}(c_0 + \frac{\log(n)}{\varepsilon\sqrt{n}})$ Chap. 4	✓	γ_k

Table 1.1 – ERM methods: Table comparing the complexity, measured in terms of iterations, of different algorithms for non-convex optimization. MC stands for Monte Carlo integration of the drift term and Step. for stepsize.

Our Contributions Besides improving the analysis of such optimization procedures in the non-convex setting, most of the existing findings hold for gradient-type algorithms. This thesis, through Chapter 4 and Chapter 5, attempts to generalize those faster rates for non-gradient, called *scaled-gradient*, type of algorithms, such as the EM method. For instance, we develop upon the MM principle several algorithms Chapters 3 and 5 for

general surrogate optimization procedure and EM type algorithms and provide their non-asymptotic analysis. A summary of our findings is given Table 2.1.

1.2.2 Stochastic Approximation

The Stochastic Approximation (SA) procedure, introduced by Robbins and Monro [1951], aims at finding a zero of a continuous function, that is only accessible through noisy evaluations. It formulates as follows:

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma_{k+1} H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) \,, \tag{1.2.13}$$

where $\boldsymbol{\theta}^{(k)} \in \Theta \subset \mathbb{R}^d$ denotes the k-th iterate, $\{\gamma_k\}_{k>0}$ is a positive deterministic sequence of stepsizes and $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ is the k-th *stochastic* update (a.k.a. drift term) depending on a random element X_{k+1} taking its values in a set X. The drift term $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ can be decomposed as a sum of a mean field h and an error term e_{k+1}

$$H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) = h(\boldsymbol{\theta}^{(k)}) + e_{k+1} \,. \tag{1.2.14}$$

In this thesis, we will focus on algorithms that use stochastic first-order oracles.

Definition 1.7 *(Stochastic first-order oracle) At iteration k the stochastic first order oracle outputs a stochastic drift term $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ where $\{X_{k+1}\}_{k>0}$ are random elements.*

Usually, the error term e_{k+1} is assumed to be an i.i.d. sequence of zero-mean finite variance noise. Formally, the following assumption is typically made:

H1.1 *The sequence of noise vectors is a Martingale difference sequence with, for any $k \in \mathbb{N}$, $\mathbb{E}\left[\boldsymbol{e}_{k+1} \,|\, \mathcal{F}_k\right] = \mathbf{0}$, $\mathbb{E}\left[\|\boldsymbol{e}_{k+1}\|^2 \,|\, \mathcal{F}_k\right] \leq \sigma_0^2 + \sigma_1^2 \|h(\boldsymbol{\theta}^{(k)})\|^2$ with $\sigma_0^2, \sigma_1^2 \in [0, \infty)$ where $\mathcal{F}_k$ denotes the filtration generated by the random variables $(\boldsymbol{\theta}^{(0)}, \{X_m\}_{m \leq k})$.*

Note that in this case, $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ is an unbiased estimator of the mean-field $h(\boldsymbol{\theta}^{(k)})$ and that the variance of $\left\| H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) - h(\boldsymbol{\theta}^{(k)}) \right\|$ is bounded.

In its original formulation (2.1.1), the population risk minimization can be performed using a SA procedure as noticed in [Bottou and Le Cun, 2005]. Particularly, stochastic gradient methods are now ubiquitous in machine learning, both from the practical side, as a simple algorithm that can learn from a single or a few passes over the data [Bottou and Le Cun, 2005], and from the theoretical side, as it leads to optimal rates for estimation problems in a variety of situations [Nemirovsky A.S. and IUdin, 1983, Polyak and Juditsky, 1992]. SA finds the minimum of the objective function by searching for roots of its gradients ($h = \nabla \mathcal{L}$) as long as it is assumed differentiable. From a machine learning point of view, this procedure access the data in a streaming fashion,*i.e.,* it can only perform a single

pass over the dataset, and minimizes the population risk which we recall is an unknown function.

Convergence of Robbins-Monro type procedures Lyapunov's second method [Kalman and Bertram, 1960] is a common method for proving the global asymptotic stability of the solutions of the Robbins-Monro procedure by showing that all the trajectories of the limiting ordinary differential equation (ODE) $\dot{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$ of this procedure go to zero. The idea is to introduce a nonnegative function which can be interpreted as an energy that decreases with each iteration of the method. In general, those Lyapunov functions are user-designed as there is no generic way to find them. In particular, we show Chapter 5 that some variants of the EM algorithm do not decrease, at each iteration, the objective function (the incomplete log-likelihood), as advocated in [Wu et al., 1983], but instead exhibit a monotonicity property of a well-designed Lyapunov function. The relation between the objective of the Robbins-Monro procedure,*i.e.,* solving $h(\boldsymbol{\theta}) = 0$, and the stationarity of the Lyapunov function is discussed Lemma 10 (Section 5.3 of Chapter 5) and Proposition 5 (Section 4.3 of Chapter 4).

Prior Work Most results available as of today [see for example [Benveniste et al., 1990], [Kushner and Yin, 2003, Chapter5, Theorem 2.1] or [Borkar, 2009]] have an asymptotic flavor. The focus of these works is to establish that the stationary point of the sequence $\{\boldsymbol{\theta}^{(k)}, k \in \mathbb{N}\}$ belongs to a stable attractor of its limiting ODE $\dot{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$.

Important advances in methodology consider the case where $\{\boldsymbol{e}_k\}_{k\geq 1}$ is state-dependent Markov noise. In this setting, the random element X_{k+1} is drawn from a state-dependent Markov process. For any bounded measurable function φ and $k \in \mathbb{N}$, we have

$$\mathbb{E}\left[\varphi(X_{kn+1}) \,|\, \mathcal{F}_k\right] = P_{\boldsymbol{\theta}^{(k)}}\varphi(X_k) = \int \varphi(x) P_{\boldsymbol{\theta}^{(k)}}(X_k, \mathrm{d}x) \;,$$

where $P_{\boldsymbol{\theta}}$ is a Markov kernel on $\mathsf{X} \times \mathcal{X}$. In general, it is assumed that $\boldsymbol{\theta} \in \Theta$, $P_{\boldsymbol{\theta}}$ has a unique stationary distribution $\pi_{\boldsymbol{\theta}}$, *i.e.,* $\pi_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}}$. Such methodologies are particularly relevant in reinforcement learning such as Q-learning [Jaakkola et al., 1994], policy gradient [Baxter and Bartlett, 2001] and temporal difference learning [Bhandari et al., 2018, Dalal et al., 2018a,b, Lakshminarayanan and Szepesvari, 2018]. Yet, their analysis is, as of today, missing in the literature.

Of course, SA schemes go far beyond gradient methods. In fact, in many important applications, the drift term of the SA is *not* a noisy version of the gradient, *i.e.,* the mean field h is not the gradient of the objective function.

These last two remarks substantiate the question asked in the previous section regarding non-gradient algorithms and their global and non-asymptotic analysis in the non-convex

setting and motivate a good core of this thesis.

Our Contributions We study Chapter 4 a general SA scheme with a potentially biased and not necessarily gradient drift term under mild conditions. An interesting extension that we focus on is the consideration of a state-dependent Markov noise that we analyze using the Poisson equation. A rigorous verification of the assumptions allow us to apply our results to several examples of interest.

1.3 Maximum Likelihood Estimation in Latent Data Models

1.3.1 Latent Data Models

In this section, we formally introduce an instance of general models class called *Latent Data Models* that are used during the *modeling* phase of the learning procedure. Let Z be a subset of $\mathbb{R}^m$, μ be a σ-finite measure on the Borel σ-algebra $\mathcal{Z} = \mathcal{B}(\mathsf{Z})$ and $\{f(z, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ be a family of positive μ-integrable Borel functions on Z. Set $z \in \mathsf{Z}$. Define, for all $\boldsymbol{\theta} \in \Theta$:

$$\begin{aligned} g(y; \boldsymbol{\theta}) &:= \int_{\mathsf{Z}} f(z, y; \boldsymbol{\theta}) \mu(\mathrm{d}z) \ , \\ p(z|y; \boldsymbol{\theta}) &:= \begin{cases} \frac{f(z,y;\boldsymbol{\theta})}{g(y;\boldsymbol{\theta})} & \text{if } g(y; \boldsymbol{\theta}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1.3.1}$$

Note that $p(z|y; \boldsymbol{\theta})$ defines a probability density function with respect to μ and $\mathcal{P} = \{p(z|y; \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta; (y, z) \in \mathsf{Y} \times \mathsf{Z}\}$ a family of probability densities. We denote by $\{\mathbb{P}_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ the associated family of probability measures. Naturally, the loss function $\mathcal{L}(\boldsymbol{\theta})$ is defined for all $\boldsymbol{\theta} \in \Theta$ as follows:

$$\mathcal{L}(\boldsymbol{\theta}) := \log g(y; \boldsymbol{\theta}) \ . \tag{1.3.2}$$

Example 1.1 *We now give some examples of latent structure:*

- *In presence of missing data, y stands for the observed data and the latent variables z are the missing data.*
- *For mixed effects models, the latent variables z are the random effects and identifying the structure of the latent data mainly corresponds to the inter-individual variability among the individuals of the dataset. This setting is presented in Section 1.4 and studied Chapter 6.*

- *For mixture models, the latent variables correspond to the unknown mixture labels taking values in a discret finite set. This setting is studied Chapters 4, 5 and 7.*

Remark 1.1 *In this thesis, we are interested in an empirical approach to the Maximum Likelihood Estimation problem. We consider n independent and not necessarily identically distributed vector of observations $(y_i \in \mathsf{Y}_i, i \in [\![1, n]\!])$ where Y_i is a subset of $\mathbb{R}^{l_i}$ and latent data $(z_i \in \mathsf{Z}_i, i \in [\![1, n]\!])$. For all $\boldsymbol{\theta} \in \Theta$, we set*

$$\begin{aligned} f(z, y; \boldsymbol{\theta}) &= \prod_{i=1}^{n} f_i(z_i, y_i; \boldsymbol{\theta}) \ , \\ g(y; \boldsymbol{\theta}) &= \prod_{i=1}^{n} g_i(y_i; \boldsymbol{\theta}) \ , \\ p(z|y; \boldsymbol{\theta}) &= \prod_{i=1}^{n} p_i(z_i|y_i; \boldsymbol{\theta}) \ . \end{aligned} \tag{1.3.3}$$

Thus, the objective function (2.3.2) formulates:

$$\mathcal{L}(\boldsymbol{\theta}) := \sum_{i=1}^{n} \log g_i(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \ . \tag{1.3.4}$$

Note that in order to avoid singularities and degeneracies of the MLE as highlighted in [Fraley and Raftery, 2007], one can regularize the objective function through a prior distribution over the model parameters, see Chapter 4 for an illustrative example.

1.3.2 The EM Algorithm

A popular class of inference algorithms to minimize (2.3.2) is the Expectation-Maximization (EM) algorithm developed in the pioneering work by Dempster et al. [1977]. The EM is an iterative procedure that minimizes the function $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta})$ when its direct minimisation is difficult. Denote by $\boldsymbol{\theta}^{(k-1)}$ the current fit of the parameter at iteration k, then the k-th step of the EM algorithm might be decomposed into two steps. The E-step consists of computing the surrogate function defined for all $\boldsymbol{\theta} \in \Theta$ as :

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) := \int_{\mathsf{Z}} p(z|y; \boldsymbol{\theta}^{(k-1)}) \log f(z, y; \boldsymbol{\theta}) \mu(\mathrm{d}z) \ . \tag{1.3.5}$$

In the M-step, the value of $\boldsymbol{\theta}$ minimizing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ is calculated and is set as the new parameter estimate $\boldsymbol{\theta}^{(k)}$. These two steps are repeated until convergence. The essence of the EM algorithm is that decreasing $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ forces a decrease of the function $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta})$, see [McLachlan and Krishnan, 2007] and the references therein.

Remark 1.2 *Using the concavity of the logarithmic function and the Jensen inequality, we can show that $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ is a majorizing surrogate function for the objective $\mathcal{L}(\boldsymbol{\theta})$*

at $\boldsymbol{\theta}^{(k-1)}$. *This scheme nicely falls into the MM principle introduced in Section 1.2.1 and is exploited in [Gunawardana and Byrne, 2005]. Chapter 5 expands this remark and develops a global analysis of an incremental variant of the EM, introduced by Neal and Hinton [1998].*

Remark 1.3 *A common assumption regarding the direct applicability of the EM for latent data models (see, in particular, the discussion of Dempster et al. [1977]) is to consider that the complete model belongs to the curved exponential family,* i.e., *for all* $\boldsymbol{\theta} \in \Theta$*:*

$$\log f(z, y, \boldsymbol{\theta}) = -\psi(\boldsymbol{\theta}) + \langle \tilde{S}(z, y), \phi(\boldsymbol{\theta}) \rangle. \tag{1.3.6}$$

where $\psi : \boldsymbol{\theta} \mapsto \mathbb{R}$ *and* $\phi : \boldsymbol{\theta} \mapsto \mathbb{R}$ *are twice continuously differentiable functions of* $\boldsymbol{\theta}$ *and* $\tilde{S} : \mathsf{Z} \mapsto \mathsf{S}$ *is a statistic taking its values in a convex subset* S *of* $\mathbb{R}$. *Then, both steps of the EM formulate in terms of sufficient statistics. Note that this assumption is rather not restrictive as many models of interest in machine learning satisfy it.*

Prior Work The EM method has been the subject of considerable interest since its formalization in [Dempster et al., 1977]. Most prior works studying the convergence of EM methods consider the *asymptotic* and/or *local* behaviors to avoid making any non-convexity assumption. The global convergence to a stationary point (either a local minimum or a saddle point) of the EM method has been established by Wu et al. [1983] as an extension of prior work developed in [Dempster et al., 1977]. The global convergence is a direct consequence of the EM method to be monotone, *i.e.,* the objective function never decreases. Locally and under regularity conditions, a linear convergence rate to a stationary point has been studied in [McLachlan and Krishnan, 2007, chapters 3 and 4]. Following Remark 1.1, a natural enhancement of those methods corresponds to constructing cheaper updates at each iteration. For instance, the convergence of the Incremental EM (iEM) method was first tackled by Gunawardana and Byrne [2005] exploiting the interpretation of the method as an alternating minimization procedure under the information geometric framework developed in [Csiszár and Tusnády, 1984]. More recently, the *local but non-asymptotic convergence* of EM methods has been studied in several works. These results typically require the initializations to be within a neighborhood of an isolated stationary point and the (negated) log-likelihood function to be strongly convex locally. Such conditions are either difficult to verify in general or have been derived only for specific models; see for example [Balakrishnan et al., 2017, Wang et al., 2015a, Xu et al., 2016a] and the references therein. The local convergence of a variance reduced EM method (called sEM-VR), that is not exactly an incremental method but rather builds upon a control variate principle, has been studied in [Chen et al., 2018, Theorem 1] under a pathwise global stability condition.

Our Contributions It is thus of utmost importance to improve and analyze EM variants in order to address the two challenges mentioned at the very beginning of this Introduction, *i.e.,* the increasing amount of data points and the non-convexity of the objective. Chapter 5 of this manuscript proposes and analyzes several variance reduced and incremental versions of the EM algorithm, such as th iEM and the sEM-VR mentioned above, in order to scale to large number of points and speed up the convergence. Particularly, this chapter considers the iEM as in [Gunawardana and Byrne, 2005] and extends their result under mild assumptions, such as the non-convexity of the objective and the uniform sampling of the indices, performed independently throughout the passes over the data.

1.3.3 The SAEM Algorithm

In many situations, the expectation step of the EM algorithm (2.3.5) can be numerically involved or even intractable. To address that issue, Wei and Tanner [1990a] propose to replace the expectation by a Monte Carlo integration, leading to the so-called Monte Carlo EM (MCEM). Another option, developed in [Delyon et al., 1999a], is the Stochastic Approximation of the EM (SAEM) that develops as follows:

1. **Simulation step:** Draw the Monte Carlo batch of latent variables $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ from its posterior distribution $p(z|y;\boldsymbol{\theta}^{(k-1)})$.

2. **Stochastic approximation step:** update the approximation, denoted $Q_k(\boldsymbol{\theta})$, of the conditional expectation (2.3.5):

$$Q_k(\boldsymbol{\theta}) = Q_{k-1}(\boldsymbol{\theta}) + \gamma_k \left(M_{(k)}^{-1} \sum_{m=1}^{M_{(k)}} \log f(z_m^{(k)}, y; \boldsymbol{\theta}) - Q_{k-1}(\boldsymbol{\theta}) \right) , \tag{1.3.7}$$

   where $\{\gamma_k\}_{k>0}$ is a sequence of decreasing stepsizes with $\gamma_1 = 1$.

3. **Maximization step:**

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q_k(\boldsymbol{\theta}) . \tag{1.3.8}$$

During the stochastic approximation phase, the conditional distribution of the parameters is obtained as it is the distribution in which the latent variables z are imputed to obtain a complete dataset from which the conditional log-likelihood is derived, see [Kuhn and Lavielle, 2004].

In the simulation step, since the relation between the observed data and the latent data can be non linear, sampling from the posterior distribution is hard and often requires using an inference algorithm. Kuhn and Lavielle [2004] proved almost-sure convergence of the sequence of parameters obtained by this algorithm coupled with an MCMC procedure during the simulation step. Indeed, $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ is a Monte Carlo batch. In simple scenar-

ios, the samples $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ are conditionally independent and identically distributed with distribution $p(z|y, \boldsymbol{\theta}^{(k-1)})$. Nevertheless, in most cases, sampling exactly from this distribution is not an option and the Monte Carlo batch is sampled by Monte Carlo Markov Chains (MCMC) algorithm.

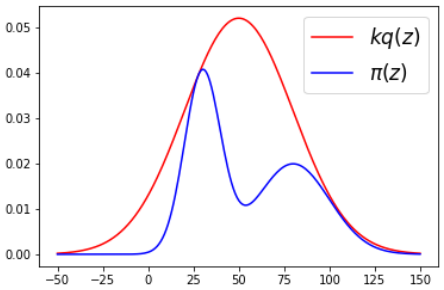

Figure 1.2 – Metropolis-Hastings (MH) algorithm: representation of a proposal $q(z)$ and the target $\pi(z)$ distributions in one dimension.

MCMC algorithms are a class of methods allowing to sample from complex distribution over (possibly) large dimensional space. An important class of samplers, called Metropolis-Hastings (MH) algorithm, iteratively draw samples from a proposal distribution q with the distribution of the newly drawn sample only depending on the current one. With some probability, the sample is either accepted as the new state of the chain or rejected. It is well-known, see [Mengersen and Tweedie, 1996, Roberts and Rosenthal, 2011] that the Independent Sampler, a sub-class of MH samplers where the proposal is independent of the current state of the chain, is geometrically ergodic if and only if, for a given ε, $\inf_{z\in\mathbb{R}^p} q(z)/\pi(z) \geq \varepsilon > 0$ where $\pi(z)$ is the target distribution. More generally, it is shown in [Roberts and Rosenthal, 2011] that the mixing rate in total variation depends on the expectation of the acceptance ratio under the proposal distribution which is also directly related to the ratio of proposal to the target. This observation naturally suggests to find a proposal which approximates the target. Figure 2.2 illustrates that remark where the proposal is a simple Gaussian distribution. From this figure, one can acknowledge that the efficacy of the sampler will be impacted by the level of similarity (*eg.* they belong to the same family of distributions) between the two distributions of interest.

In the stochastic approximation step, the sequence of decreasing positive integers $\{\gamma_k\}_{k>0}$ controls the convergence of the algorithm. In practice, γ_k is set equal to 1 during the first K_1 iterations to let the algorithm explore the parameter space without memory and to converge quickly to a neighborhood of the ML estimate. The stochastic approximation is performed during the final K_2 iterations where $\gamma_k = 1/k^a$ with in general $a = 0.7$, ensuring the almost sure convergence of the estimate.

Prior Work The SAEM algorithm has been shown theoretically to converge to a maximum of the likelihood of the observations under very general conditions [Delyon et al.,

1999a]. As already mentioned, this result has been extended by Kuhn and Lavielle [2004], to include an MCMC sampling scheme in the simulation phase. Recent work by Allassonnière and Chevallier [2019], exhibits a new class of algorithms where the simulation step is performed using an annealed version of the posterior distribution and is motivated by saddle points escaping problems.

Our Contributions We consider the SAEM algorithm through Chapters 6 and 7. In particular Chapter 6 contains an improvement of the aforementioned sampling procedure. We exploit the remark about the Independent Sampler by introducing an efficient MH proposal based on the Laplace approximation of the incomplete log likelihood. A linearisation of the structural model is shown to be equivalent for the class of continuous data models. Chapter 7 exploits the finite sum structure of the objective function (following Remark 1.1) and proposes an incremental variant of the SAEM which asymptotic behavior is shown theoretically and experimentally.

1.4 Mixed Effects Modeling and Population Approach

1.4.1 Why Are Mixed Effects Models Relevant?

Mixed Effects Models (MEM), see [Lavielle, 2014] and the references therein, have received increasing use due to their flexibility for analyzing multi-outcome longitudinal data following possibly nonlinear profiles. They are reference methods to describe inter-individual variabilities among a given population.

A general formulation of the MEM for the continuous observation y_{ij} can be written as follows:

$$y_{ij} = \mathsf{f}(x_{ij}, \psi_i) + \mathsf{g}(x_{ij}, \psi_i, \xi_i)\, \varepsilon_{ij} \quad \text{with} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)\ , \tag{1.4.1}$$

where the quantity y_{ij} denotes the j-th observation for the i-th individual. x_{ij} is a vector of regressors (it could be the time or some variables such as the time or the dose of a drug injected), f is the (possibly nonlinear) structural model, ψ_i denotes the individual parameters. The quantity ε_{ij} is a random variable assumed to be normally distributed and σ denotes the variance parameter entering the function g, which expresses the standard deviation of the measurement error and is generally either constant (homoscedastic variance) or a function of f.

We consider here a two-stage model, as in [Davidian, 2017], which both provides a typical population curve, also known as the *structural model* (see the function f in (2.4.1)), and models the individual parameters, denoted ψ_i and regarded as random variables that fluctuate around a population parameter ψ_{pop}. This latter probabilistic model of the

individual parameters exhibits the inter-individual variability structure that governs the statistical phenomena. Formally, it reads:

$$\begin{cases} \psi_i = \mathsf{h}(\psi_{\text{pop}}, \eta_i) \\ \eta_i \sim \mathcal{N}(0, \Omega) \end{cases} \tag{1.4.2}$$

We note that the individual parameters ψ_i are related through a function h to ψ_{pop}, the p-dimensional vector containing the fixed effects, and η_i, the q-dimensional vector containing the random effects η_i. For instance, for Normal individual parameters we have $\psi_i = \psi_{\text{pop}} + \eta_i$ and for Lognormal we have $\psi_i = \psi_{\text{pop}}\, \mathrm{e}^{\eta_i}$. The random effects η_i and the residual errors ε_{ij} are assumed to be independent for different subjects and to be independent of each other for the same subject.

The objective here is to estimate the vector of parameters $\boldsymbol{\theta} = (\psi_{\text{pop}}, \Omega, \sigma^2)$ by maximum likelihood. In mixed effects models, the likelihood associated with (2.4.1) and (2.4.2) is intractable as individual likelihoods need to integrate out the unknown parameters ψ_i over their distribution.

For MEMs, the expectation computed in (2.3.5) is intractable due to the possible nonlinearity of the structural model. We thus use the SAEM algorithm introduced above where the latent variables, that are simulated at each iteration, correspond to the individual parameters ψ_i.

1.4.2 Application to Population Pharmacokinetics

In domains such as economy, sociology, genomics or pharmacokinetics-pharmacodynamics (PK-PD), observations from several individuals of a population are measured. Consider the observations Figure 2.3.

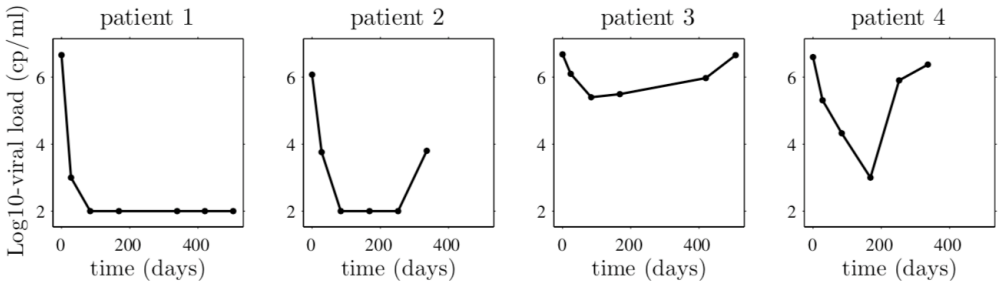

Figure 1.3 – Viral load of four patients with hepatitis C (taken from [Lavielle, 2014]).

These measurements are viral loads for four different patients with hepatitis C (HCV) after treatment that started at time $t = 0$. We note that in many cases, such as the one

exposed here, data are *longitudinal*, *i.e.,* they are repeated measurements (not necessarily function of time) of an output quantity. Also, every patient does not react to the treatment the same way. For instance, patient 1 sees its viral load decrease after the treatment while patient 3 has almost no response. Following those two remarks, the best way to cope with statistical modeling of the HCV dynamics is to build a model that describes how the viral load evolves with time and a statistical model that explains the difference among patients. Mixed Effects Modeling is a natural framework for such analysis and is thoroughly developed in [Snoeck et al., 2010] for HCV dynamics modeling.

The so-called *population approach* becomes very relevant in this cas as it combines (possibly) poor individual information to build a comprehensive population model.

MEMs and the SAEM algorithm are extensively used to handle such experiments through their implementation in software tools such as Monolix, NONMEM, the SAEMIX R package [Comets et al., 2017] and the nlmefitsa Matlab function. Part of our work in this thesis relies on the SAEMIX Package (R [R Development Core Team, 2008]), see Chapters 6-8.

Our Contributions Several PK models are studied through Chapters 6-8 using the Mixed Effects Modeling and the population approach. We apply and show the efficacy of our newly developed methods to accelerate the MLE phase. An extension of the SAEMIX R package for noncontinuous data models is also presented Chapter 8.

Chapter 2

Introduction en Français

Abstract: *Ce chapitre introductif décrit les objectifs de la thèse et introduit les principaux domaines étudiés dans les chapitres qui suivent. Nous donnons, ici, une vision approfondie de la literature en lien avec ces domaines et insistons sur le gap que ce manuscrit essaye de combler. D'importantes hypothèses et définitions, faites tout au long de la thèse, sont présentées dans ce chapitre afin de se familiariser avec l'optimisation non-convexe, l'approximation stochastique et les modèles à données latentes. La dernière section développe un example spécifique des modèles à données latentes appelé modèles à effets mixtes ainsi que son application à la pharmacologie, comme domaine d'intérêt de notre équipe XPOP, INRIA.*

Contents

2.1 Apprentissage Statistique

Le domaine de la modélisation mathématique a été au coeur de l'effort humain destiné à mieux comprendre le monde, avec des applications allant de la physique aux sciences sociales. En particulier, pour traiter un grand nombre de données et modéliser des phénomènes complexes, l'apprentissage statistique est considéré comme l'un des sous-domaines les plus importants de notre époque. Il peut être considéré comme une approche fondée sur des principes d'extraction d'informations utiles à partir de données qui peuvent être exploitées pour exécuter des tâches telles que la prédiction. Il s'agit généralement d'une phase de *modélisation*, où un modèle est conçue dans un espace de recherche de modèles donné — dans cette thèse, nous nous limitons aux modèles paramétriques où l'espace de recherche est un ensemble de paramètres — et d'une phase *d'entrainement* ou *d'optimisation* où, pour des paires d'observations entrées-sorties, le modèle est adapté pour décrire au mieux les données. Nous donnons maintenant une formulation rigoureuse des idées présentées ci-dessus.

Formulation mathématique Considérons la paire de variables aléatoires entrées-sorties (X, Y) prenant des valeurs dans un ensemble d'entrées arbitraires $\mathsf{X} \subset \mathbb{R}^p$ et un ensemble de sorties arbitraires $\mathsf{Y} \subset \mathbb{R}^q$. Par exemple, X est une matrice de covariables décrivant un patient hospitalisé (âge, poids, etc.) et Y décrit sa charge virale pour l'hépatite C. Nous désignons par $\mathcal{P}$, la distribution selon laquelle cette paire entrée-sortie est tirée. Comme mentionné plus haut, la phase de *modélisation* consiste à trouver une fonction mesurable $M_{\boldsymbol{\theta}} : \mathsf{X} \mapsto \mathsf{Y}$ qui est dans notre cas une fonction paramétrique de paramètre $\boldsymbol{\theta} \in \mathbb{R}^d$. Cette fonction est communément appelée le *prédicteur* et sa performance est mesurée par une fonction de *coût* $\ell : \mathsf{Y} \mapsto \mathbb{R}$ où $\ell(y, y')$ est la perte subie quand la vraie sortie est y alors que y' est prédit. Ensuite, la phase *d'entrainement* se résume à calculer la quantité suivante :

$$\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) = \arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \{\mathcal{L}(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta})\} \quad \text{avec} \quad \mathcal{L}(\boldsymbol{\theta}) = \mathbf{E}_{(x,y)\sim\mathcal{P}}\left[\ell(y, M_{\boldsymbol{\theta}}(x))\right] , \quad (2.1.1)$$

où ℓ est une fonction de perte éventuellement non convexe et fonction de données observées, $\mathcal{L}$ est ce qu'on appelle le *risque de population* et $\mathrm{R}(\cdot)$ est un terme de pénalisation qui impose une structure à la solution et est éventuellement non lisse.

Tout au long de cette thèse, nous nous intéressons aux modèles où la relation entrée-sortie n'est pas complètement caractérisée par les paires $(x, y) \in \mathsf{X} \times \mathsf{Y}$ observées, mais dépend aussi d'un ensemble de variables latentes non observées $z \in \mathsf{Z} \subset \mathbb{R}^m$. Ces modèles sont appelés modèles à données latentes et sont formellement introduits Section 1.3. Ils incluent le cadre des données incomplètes, *i.e.,* certaines observations manquent, mais est beaucoup plus large que cela (par exemple, la structure latente peut correspondre aux

labels inconnues dans les modèles de mélange ou aux états cachés dans les modèles de Markov cachés). Dans tous ces cas, une étape de simulation est nécessaire pour compléter les données observées par des réalisations des variables latentes. Cette dernière étape de simulation joue un rôle clé dans ce manuscrit et est traitée en détail dans chaque chapitre. Formellement, cette spécificité dans notre contexte implique d'étendre la fonction de perte ℓ pour accepter un troisième argument comme suit :

$$\ell(y, M_{\boldsymbol{\theta}}(x)) = \int_{\mathsf{Z}} \ell(z, y, M_{\boldsymbol{\theta}}(x)) \mathrm{d}z \ . \tag{2.1.2}$$

Notez que, pour des raisons de notation, nous utilisons le même nom pour les deux fonctions de perte définies sur des espaces différents. Enfin, nous considérons des exemples où la fonction $\mathcal{L}$ est lisse dans le sens suivant :

Definition 2.1 *Une fonction $f : \mathbb{R}^d \mapsto \mathbb{R}$ est L-smooth si et seulement si elle est différentiable et son gradient est L-Lipschitz-continu,* i.e., *pour tout $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$:*

$$\|\nabla f(\theta) - \nabla f(\vartheta)\| \leq L \|\theta - \vartheta\| \ . \tag{2.1.3}$$

Traditionnellement, l'apprentissage statistique s'est surtout concentré sur le développement de fonctions de perte convexe ℓ et d'algorithmes tels que SVM ou des modèles graphiques à famille exponentielle. Cependant, de nombreux problèmes importants, tels que la vision par ordinateur et le traitement du langage naturel, ne peuvent être formulés comme une optimisation convexe ou, en tout cas, seront plus coûteux en termes de calcul que leurs équivalents non convexes. En effet, si la convexité peut être considérée comme une vertu, elle peut aussi être considérée comme une limitation dans la complexité du modèle choisi pour résoudre un problème donné. Par exemple, les modèles à variables latentes, mentionnés plus haut comme une grande famille de modèles graphiques probabilistes, impliquent une optimisation non convexe et sont utiles pour s'attaquer à des tâches telles que la reconnaissance vocale (réalisée par exemple avec des modèles de mélanges gaussiens), qui ne peuvent être traitées avec un modèle convexe.

L'augmentation de la dimension/taille de l'échantillon et la complexité des tâches obligent la communauté des statisticiens à développer des algorithmes plus simples, avec une complexité maximale de $\mathcal{O}(n)$ où n est soit la dimension soit le nombre d'observations, tout en s'adaptant à des modèles plus complexes et fortement non convexes. Cette question est traitée en détail dans [Bottou and Bousquet, 2008] et est à l'origine de l'expansion du domaine de l'optimisation non convexe.

2.2 Optimisation Non-convexe

Les problèmes d'optimisation non convexe sont fréquents dans l'apprentissage machine, comme dans la sélection des caractéristiques, l'apprentissage matriciel structuré, l es modèles de mélanges et l'entrainement des réseaux de neurones. Dans tous ces cas, la fonction $\mathcal{L}$ définie dans l'objectif d'optimisation (2.1.1) est non convexe. La convexité d'une fonction f est définie comme suit

Definition 2.2 *Une fonction $f : \mathbb{R}^d \mapsto \mathbb{R}$ est dite convexe si pour tout $(\theta, \vartheta) \in \mathbb{R}^d \times \mathbb{R}^d$ et tout $\lambda \in (0, 1)$:*

$$f((1-\lambda)\theta + \lambda\vartheta) \leq (1-\lambda)f(\theta) + \lambda f(\vartheta) \; . \tag{2.2.1}$$

Dans ce manuscrit, nous nous intéressons à la formulation sous contrainte de ce problème d'optimisation. Ainsi, le vecteur de paramètres $\boldsymbol{\theta}$ appartient à un ensemble convexe $\Theta \subset \mathbb{R}^d$ dans le sens suivant :

Definition 2.3 *Un ensemble Θ est dit convexe si pour tout $(\theta, \vartheta) \in \Theta^2$ et tout $\lambda \in (0, 1)$:*

$$(1-\lambda)\theta + \lambda\vartheta \in \Theta \; . \tag{2.2.2}$$

La différentiabilité de la fonction de l'objectif sur un ensemble contraint est traitée en introduisant le concept suivant de différentiabilité directionnelle (qui inclut la notion de différentiabilité) :

Definition 2.4 *Pour toute fonction $f : \Theta \to \mathbb{R}$, $f'(\boldsymbol{\theta}, \boldsymbol{d})$ est le dérivé directionnel de f à $\boldsymbol{\theta}$ suivant la direction $\boldsymbol{d}$,* i.e.,

$$f'(\boldsymbol{\theta}, \boldsymbol{d}) := \lim_{t \to 0^+} \frac{f(\boldsymbol{\theta} + t\boldsymbol{d})) - f(\boldsymbol{\theta})}{t} \; . \tag{2.2.3}$$

Analyser de la convergence de l'algorithme d'optimisation, dite "convexe". (resp. *non-convexe*) si l'objectif (2.1.1) est *convexe* (resp. *non-convexe*), implique généralement une condition de sous-optimalité comme critère de convergence. Par exemple, pour les fonctions convexes, nous utilisons $|\mathcal{L}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}^*)|$ (ou $\|\boldsymbol{\theta} - \boldsymbol{\theta}^*\|^2$) comme condition. Nous désignons par $\boldsymbol{\theta}^*$ la solution optimale que l'on peut trouver efficacement dans le cas convexe. Par conséquent, lorsqu'il est difficile de trouver une telle solution optimale, comme dans le cas non convexe, ce critère de convergence ne peut pas tenir. Nous utilisons alors la quantité $\|\nabla\mathcal{L}(\boldsymbol{\theta})\|^2$, comme préconisé dans [Ghadimi and Lan, 2013] et [Nesterov, 2004], pour évaluer la stationnarité des resultants de l'algorithme. La définition suivante est donc importante tout au long de notre analyse :

Definition 2.5 *Un point $\boldsymbol{\theta}^*$ est dit être ε-stationnaire si $\|\nabla\mathcal{L}(\boldsymbol{\theta}^*)\|^2 \leq \varepsilon$. Un algorithme itératif stochastique est dit d'atteindre ε-stationnarité en $R > 0$ itérations si $\mathbb{E}[\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(R)})\right\|^2] \leq \varepsilon$, où l'espérance est prise est sur la stochasticité de l'algorithme.*

Nous donnons deux formulations, trouvées dans la littérature, de tels résultats dans le cas convexe et non convexe pour donner une idée du type de bornes que l'on peut obtenir pour caractériser la stationnarité des itérations de l'algorithme. Considérons le problème d'optimisation simple suivant, non contraint et non régularisé :

$$\arg\min_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \ , \tag{2.2.4}$$

où $\mathcal{L} : \mathbb{R}^d \mapsto \mathbb{R}$ is L-smooth. Dans le cas convexe, un résultat établi est:

Proposition 3 *(Convergence de la descente de gradient [Nesterov, 2004] pour les fonctions convexes). Considérons le schéma simple de descente de gradient, avec pas constant, qui commence à partir d'un $\boldsymbol{\theta}^{(0)}$ initial et et dont la séquence des itérations $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ s'exprime comme suit :*

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma \nabla \mathcal{L}(\boldsymbol{\theta}^{(k)}) \ , \tag{2.2.5}$$

où $\mathcal{L}$ est une fonction convexe et L-smooth sur $\mathbb{R}^d$. Soit le pas $\gamma = 1/L$, alors la séquence d'itérations $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ satisfait :

$$\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\mathcal{L}(\boldsymbol{\theta}^*) \leq \frac{L \left\| \boldsymbol{\theta}^{(k+1)} - \boldsymbol{\theta}^{(0)} \right\|^2}{k+1} \ . \tag{2.2.6}$$

De plus, si $\mathcal{L}$ est μ- fortement convexe nous avons

$$\mathcal{L}(\mathcal{L}(\boldsymbol{\theta}^{(k+1)}) - \mathcal{L}(\boldsymbol{\theta}^*) \leq (1 - \mu/L)^{k+1} [\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^*)] \tag{2.2.7}$$

Dans le cas non convexe, un résultat établi est:.

Proposition 4 *(Convergence de la descente de gradient stochastique [Ghadimi and Lan, 2013] pour les fonctions non convexes). Considérons la valeur initiale $\boldsymbol{\theta}^{(0)}$, un point de terminaison Γ tiré selon une fonction de masse de probabilité $P_\Gamma(\cdot)$ supportée sur $\{1, \cdots, K\}$ avec K par itération limite et les mises à jour suivantes pour $k \in [\![1, \Gamma]\!]$:*

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma_{k+1} H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) \ , \tag{2.2.8}$$

où $\{X_{k+1}\}_{k<\Gamma}$ sont i.i.d., de moyenne nulle et $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ est une estimation non biaisée du gradient $\nabla\mathcal{L}(\boldsymbol{\theta}^{(k)})$ et $\mathcal{L}$ est L-smooth et une fonction (éventuellement) non convexe sur $\mathbb{R}^d$. Supposons que

$$\mathbb{E}[H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})] = \nabla\mathcal{L}(\boldsymbol{\theta}^{(k)}) \quad \textit{and} \quad \mathbb{E}[\left\| H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) - \nabla\mathcal{L}(\boldsymbol{\theta}^{(k)}) \right\|^2] \leq \sigma^2 \tag{2.2.9}$$

et que le pas $\gamma_k < 1/2L$*, alors la séquence des itérations* $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ *satisfait :*

$$\frac{1}{L}\mathbb{E}[\left\|\nabla\mathcal{L}(\boldsymbol{\theta}^{(\Gamma)})\right\|^2] \leq \frac{D_{\mathcal{L}}^2 + \sigma^2\sum_{k=1}^K \gamma_k^2}{\sum_{k=1} K(2\gamma_k - L\gamma_k^2)} \tag{2.2.10}$$

avec $D_{\mathcal{L}} = \sqrt{2(\mathcal{L}(\mathcal{L}(\boldsymbol{\theta}^{(0)}) - \mathcal{L}(\boldsymbol{\theta}^*)/L}$*. L'espérance est prise sur la stochasticité de l'algorithme.*

En particulier, nous étendons dans le Chapitre 4 le résultat ci-dessus pour un algorithme non-gradient et lorsque le terme de dérive est un estimateur biaisé du champ moyen.

2.2.1 Minimisation du Risque Empirique

En général, comme la distribution $\mathcal{P}$ générant les données est souvent inconnue, n paires $((y_i, x_i), i \in [\![1, n]\!])$ d'observations, aussi appelées *examples d'entraînement*, sont considérées dans la procédure d'optimisation (2.1.1). Basé sur le principe de minimisation du risque empirique (MRE) [Vapnik, 2013], les problèmes d'optimisation impliquent une fonction de perte $\mathcal{L}$, également connue sous le nom de *risque empirique*, moyennée sur les points d'observations. Alors, la fonction objective, sans pénalisation, se lit :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) &= n^{-1}\sum_{i=1}^n \ell(y_i, M_{\boldsymbol{\theta}}(x_i)) \\ &= n^{-1}\sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}) \ .\end{aligned} \tag{2.2.11}$$

où n est le nombre d'observations, et $\mathcal{L}_i$ est la perte associée à la i-ème observation. La variante de la MRE régularisée consiste à ajouter un régularisateur éventuellement non lisse $\mathrm{R}(\boldsymbol{\theta})$ tel qu'introduit dans (2.1.1). Par exemple pour une observation $y \in \mathsf{Y}$ et une prédiction $y' \in \mathsf{Y}$ les pertes habituelles sont, la perte quadratique $\ell(y, y') = \|y - y'\|^2/2$ pour une tâche de régression et la perte logistique $\ell(y, y') = \log(1 + \exp(-\langle y \,|\, y'\rangle))$ pour la tâche de classification où nous rappelons que la prédiction y' dépend des covariables observées x, un modèle $M_{\boldsymbol{\theta}}(\cdot)$ et éventuellement de variables latentes z.

Dans le cas convexe, de nombreuses méthodes déterministes bien connues telles que la descente de gradient, les méthodes de gradient accéléré et les méthodes de Newton sont utilisées pour effectuer la tâche d'optimisation, voir [Bertsekas, 1999, Boyd and Vandenberghe, 2004, Nesterov, 2004] et les références qui y figurent. Cependant, chacune de ces méthodes est coûteuse en termes de calcul puisqu'elle nécessite un passage complet sur l'ensemble de données à chaque itération. Pour traiter un grand nombre de points n, les méthodes d'optimisation stochastique et incrémentale de premier et de second ordre

sont populaires et largement étudiées lorsque l'objectif est convexe, voir [Defazio et al., 2014, Mairal, 2015b, Roux et al., 2012, Vanli et al., 2018]. Par exemple les algorithmes incrémentaux, affichant un coût par itération moins élevé, au prix d'un plus grand coût mémoire. Pour les fonctions objectives non convexes, des méthodes déterministes [Agarwal et al., 2017, Carmon et al., 2017] et stochastiques [Allen-Zhu and Hazan, 2016, J. Reddi et al., 2016] ont également été développées pour atteindre un point ε-stationnaire. Il est important de noter que dans le cas non convexe, un point ε-stationnaire peut être un point selle. De nombreux travaux importants ont été réalisé dans l'optique d'échapper à ces points selle pour atteindre un minimum local de (2.1.1), comme dans [Reddi et al., 2018, Royer and Wright, 2018, Xu et al., 2018], mais sont hors du cadre de cette thèse.

Une classe d'algorithmes populaire pour résoudre la minimisation d'une fonction composite non convexe est celle des techniques de majorisation-minimisation [Lange, 2016] qui approchent de façon itérative de la fonction composite non convexe par une fonction de majorisation facile à minimiser. Par exemple, la plupart des techniques, comme la descente en gradient, utilisent un majorant convexe quadratique qui peut être optimisé efficacement. Une illustration de ce concept est fournie Figure 2.1.

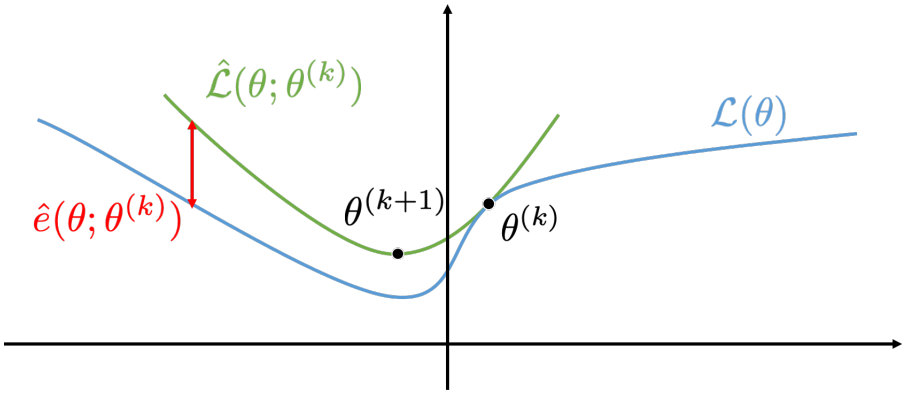

Figure 2.1 – Principe de Majorisation-Minimisation.

Notons qu'à l'itération k, l'objectif $\mathcal{L}(\boldsymbol{\theta})$ est borné par une fonction de substitut $\hat{\mathcal{L}}(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k)})$ au point d'estimation actuel $\boldsymbol{\theta}^{(k)}$. Une variante incrémentale exploitant la structure en somme finie du problème a été développée par [Mairal, 2015b] et est étendue Chapitre 3. En particulier, notre extension s'appuie sur un schéma doublement stochastique : le premier niveau de stochasticité découle du choix de l'indice i, exploitant ainsi la structure en somme finie du problème, tandis que le second niveau de stochasticité découle de la structure latente du problème utile à la construction de fonctions de substitut appropriées. Un exemple bien connu de ce cadre incrémental est l'algorithme EM (Expectation Maximization) incrémental développé dans l'ouvrage pionnier par Neal and Hinton [1998]. Les auteurs profitent de la structure latente du problème, présentée en Section précédente, pour construire des fonctions majorants *facile à optimiser* (voir Chapitre 5 pour une compréhension complète de cette approche). Dans cette thèse, nous nous concentrerons sur

les algorithmes qui utilisent des oracles incrémentaux de premier ordre.

Definition 2.6 *(Oracle incrémental de premier ordre) Pour une fonction donnée $f : \Theta \mapsto \mathbb{R}$ avec une structure en somme finie, un oracle incrémental de premier ordre prend un index $i \in [\![1, n]\!]$ et un paramètre $\boldsymbol{\theta} \in \Theta$ et renvoie les valeurs de $f_i(\boldsymbol{\theta})$ et/ou son gradient $\nabla f_i(\boldsymbol{\theta})$.*

Dans l'état actuel de la littérature, ces algorithmes sont privilégiés car ils ne nécessitent qu'une petite quantité d'informations de premier ordre à chaque itération.

Travaux antérieurs Dans le cas (possiblement fortement) convexe, la descente du gradient stochastique (SGD) a été au centre d'énormes progrès au cours de cette dernière décennie. De nombreuses variantes incrémentales [Bertsekas, 2011] ont été développées depuis son introduction dans le travail fondateur [Robbins and Monro, 1951]. Parmi eux, il est prouvé qu'une classe d'algorithmes à variance réduite permet d'atteindre des vitesses plus rapides pour des objectifs convexes. Par exemple [Defazio et al., 2014, Roux et al., 2012] développe des algorithmes incrémentaux rapides qui permettent d'obtenir des taux de convergence linéaire pour des fonctions fortement convexes. La méthode SVRG [Johnson and Zhang, 2013] est une autre méthode à variance réduite qui affiche un besoin de stockage inférieur à ces dernières. De plus, une étude des limites inférieures pour un problème d'optimisation de fonction composite a été faite dans [Agarwal and Bottou, 2014], mais la littérature reste plutôt pauvre pour un fonction non convexe.

Dans le cas non convexe, plusieurs travaux importants [Bottou, 1991, Kushner and Clark, 2012] développent une convergence asymptotique des variantes incrémentales de SGD vers un point fixe. Le premier taux de convergence non asymptotique de SGD dans [Ghadimi and Lan, 2013] assure un point stationnaire ε en $\mathcal{O}(1/\varepsilon^2)$ itérations, voir Tableau **??**. Les variantes incrémentales sont également analysées dans [Ghadimi et al., 2016] et en particulier le SVRG est connu pour atteindre un point ε-stationnaire en $\mathcal{O}(n^{2/3}/\varepsilon)$ itérations, voir [Reddi et al., 2016a]. Ces résultats sont pertinents en ce sens qu'ils se distinguent d'une hypothèse de convexité locale et abordent le comportement de convergence globale des méthodes d'optimisation dans un cadre non convexe. Ce manuscrit suit le même esprit.

En plus d'améliorer l'analyse de ces procédures d'optimisation dans un cadre non convexe, la plupart des résultats existants s'appliquent aux algorithmes de type gradient. Cette thèse, à travers les Chapitre 4 et Chapitre 5, tente de généraliser ces taux de convergence pour les algorithmes de type non gradient, tels que la méthode Expectation-Maximization (EM).

Algorithm	Gradient	Non-gradient	MC	Step.
SGD	$\mathcal{O}(1/\varepsilon^2)$ [Ghadimi and Lan, 2013]	?	×	γ_k
GD	$\mathcal{O}(n/\varepsilon)$ [Nesterov, 2004]	?	×	γ
SVRG/SAGA	$\mathcal{O}(n^{2/3}/\varepsilon)$ [Reddi et al., 2016b]	$\mathcal{O}(n^{2/3}/\varepsilon)$ Chap. 5	×	γ_k
MISO	$\mathcal{O}(n/\varepsilon)$ Chap. 3	$\mathcal{O}(n/\varepsilon)$ Chap. 3	×	—
MISSO	$\mathcal{O}(n/\varepsilon)$ Chap. 3	$\mathcal{O}(n/\varepsilon)$ Chap. 3	✓	—
Biased SA	$\mathcal{O}(c_0 + \frac{\log(n)}{\varepsilon\sqrt{n}})$ Chap. 4	$\mathcal{O}(c_0 + \frac{\log(n)}{\varepsilon\sqrt{n}})$ Chap. 4	✓	γ_k

Table 2.1 – Méthodes de MRE: Tableau de comparaison de complexité, mesuré en termes d'iterations, de différents algorithmes d'optimisation non convexe. MC siginifie Intégration de Monte Carlo du terme de dérive.

2.2.2 Approximation Stochastique

La procédure d'approximation stochastique (SA), introduite par Robbins and Monro [1951], vise à trouver un zéro d'une fonction continue, qui n'est accessible que par ses évaluations aléatoires. Sa formulation est la suivante :

$$\boldsymbol{\theta}^{(k+1)} = \boldsymbol{\theta}^{(k)} - \gamma_{k+1} H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) \,, \tag{2.2.12}$$

où $\boldsymbol{\theta}^{(k)} \in \Theta \subset \mathbb{R}^d$ dénote la k-eme itération, $\{\gamma_k\}_{k>0}$est une séquence de pas positifs et $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ est la k-eme mis à jour *stochastique* qui dpend de l'élément alátoire X_{k+1} à valeur dans X. Le terme de dérive $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$peut être décomposé en somme d'un champ moyen h et un terme d'erreur e_{k+1}

$$H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) = h(\boldsymbol{\theta}^{(k)}) + e_{k+1} \,. \tag{2.2.13}$$

Dans cette thèse, nous nous concentrerons sur les algorithmes qui utilisent des oracles stochastiques de premier ordre.

Definition 2.7 *(Oracle stochastique de premier ordre) A l'itération k l'oracle stochastique de premier ordre produit un terme de dérive stochastique $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ était $\{X_{k+1}\}_{k>0}$ sont des éléments aléatoires.*

Habituellement, le terme d'erreur e_{k+1} est supposé être une séquence i.i.d. de bruit de variance finie et de moyenne nulle. Formellement, l'hypothèse suivante est généralement formulée :

H2.1 *La séquence de bruits est un incrément Martingale avec, pour tout $k \in \mathbb{N}$, $\mathbb{E}\left[\boldsymbol{e}_{k+1} \,|\, \mathcal{F}_k\right] = \mathbf{0}$, $\mathbb{E}\left[\|\boldsymbol{e}_{k+1}\|^2 \,|\, \mathcal{F}_k\right] \leq \sigma_0^2 + \sigma_1^2 \|h(\boldsymbol{\theta}^{(k)})\|^2$ et $\sigma_0^2, \sigma_1^2 \in [0, \infty)$ où $\mathcal{F}_k$ désigne la filtration générée par les variables aléatoires $(\boldsymbol{\theta}^{(0)}, \{X_m\}_{m \leq k})$.*

Notez que dans ce cas, $H_{\boldsymbol{\theta}^{(k)}}(X_{k+1})$ est un estimateur non biaisé du champ moyen $h(\boldsymbol{\theta}^{(k)})$ et que la variance de $\left\| H_{\boldsymbol{\theta}^{(k)}}(X_{k+1}) - h(\boldsymbol{\theta}^{(k)}) \right\|$ est borné.

Dans sa formulation originale (2.1.1), la minimisation du risque de population peut être réalisée à l'aide d'une procédure SA comme indiqué dans [Bottou and Le Cun, 2005]. En particulier, les méthodes de gradient stochastique sont maintenant omniprésentes dans l'apprentissage machine, tant du point de vue pratique, en tant qu'algorithme simple qui peut apprendre d'un seul ou de quelques passages sur les données [Bottou and Le Cun, 2005], que du point de vue théorique, car il conduit à des taux optimaux pour des problèmes d'estimation dans diverses situations [Nemirovsky A.S. and IUdin, 1983, Polyak and Juditsky, 1992]. SA trouve le minimum de la fonction objective en recherchant les racines de son gradient ($h = \nabla\mathcal{L}$) tant qu'il est supposé différentiable. Du point de vue de l'apprentissage machine, cette procédure accède aux données en continu, c'est-à-dire qu'elle ne peut effectuer qu'un seul passage sur l'ensemble de données, et minimise le risque de population qui, on le rappelle, est une fonction inconnue.

Convergence des procédures de type Robbins-Monro La deuxième méthode de Lyapunov [Kalman and Bertram, 1960] est une méthode courante pour prouver la stabilité asymptotique globale des solutions de la procédure Robbins-Monro en montrant que toutes les trajectoires de l'équation différentielle ordinaire limite (EDO) $\dot{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$ de cette procédure passent par zéro. L'idée est d'introduire une fonction non négative, généralement désignée par V, qui peut être interprétée comme une énergie qui diminue à chaque itération de la méthode. En général, ces fonctions de Lyapunov sont conçues par l'utilisateur car il n'existe aucun moyen générique de les trouver. En particulier, nous montrons Chapitre 5, que certaines variantes de l'algorithme Expectation-Maximization (EM) ne diminuent pas, à chaque itération, la fonction objective (la log-vraisemblance incomplète), comme préconisé dans [Wu et al., 1983], mais montrent plutôt une propriété monotonique d'une fonction de Lyapunov bien conçue. La relation entre l'objectif de la procédure Robbins-Monro,*i.e.,* résoudre $h(\boldsymbol{\theta}) = 0$, et la stationnarité de la fonction Lyapunov est discutée Lemma 10 (Section refsec:main du Chapitre 5) et Proposition 5. (Section 4.3 du Chapitre 4).

Travaux antérieurs La plupart des résultats disponibles à ce jour[voir par exemple [Benveniste et al., 1990], [Kushner and Yin, 2003, Chapitre 5, Théorème 2.1] ou [Borkar, 2009]] ont une saveur asymptotique. L'objectif de ces travaux est d'établir que le point stationnaire de la séquence $\{\boldsymbol{\theta}^{(k)}, k \in \mathbb{NN}\}$ appartient à un attracteur stable de son ODE limite $\dot{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$.

Les progrès méthodologiques importants considèrent le cas où $\{\boldsymbol{e}_k\}_{k\geq 1}$ est le bruit de Markov dépendant de l'état. Dans cette option, l'élément aléatoire X_{k+1} est tiré d'un processus de Markov dépendant de l'état. Pour toute fonction mesurable limitée φ et

$k \in \mathbb{N}$, nous avons

$$\mathbb{E}\left[\varphi(X_{kn+1}) \,|\, \mathcal{F}_k\right] = P_{\boldsymbol{\theta}^{(k)}}\varphi(X_k) = \int \varphi(x) P_{\boldsymbol{\theta}^{(k)}}(X_k, \mathrm{d}x) \;,$$

où $P_{\boldsymbol{\theta}}$ est un noyau de Markov sur $\mathsf{X}\times\mathcal{X}$. En général, on suppose que pour $\boldsymbol{\theta} \in \Theta$, $P_{\boldsymbol{\theta}}$ a une distribution stationnaire unique $\pi_{\boldsymbol{\theta}}$, *i.e.,* $\pi_{\boldsymbol{\theta}} P_{\boldsymbol{\theta}} = \pi_{\boldsymbol{\theta}}$. Ces méthodologies sont particulièrement pertinentes pour l'apprentissage par renforcement tel que le Q-learning [Jaakkola et al., 1994], le gradient sur les politiques [Baxter and Bartlett, 2001] et l'apprentissage par différence temporelle [Bhandari et al., 2018, Dalal et al., 2018a,b, Lakshminarayanan and Szepesvari, 2018]. Pourtant, leur analyse est, à ce jour, absente de la littérature.

Bien entendu, les algorithmes de type SA vont bien au-delà des méthodes de gradient. En fait, dans de nombreuses applications importantes, le terme de dérive de la SA n'est pas une version bruyante du gradient, c'est-à-dire que le champ moyen h n'est pas le gradient de la fonction objective.

Ces deux dernières remarques corroborent la question posée dans la section précédente concernant les algorithmes non gradient et leur analyse globale/non-asymptotique dans le cadre non convexe et motivent une importante partie de cette thèse.

2.3 Maximum de Vraisemblance Dans Des Modèles à Données Latentes

2.3.1 Modèles à Données Latentes

Dans cette section, nous présentons formellement une instance des modèles généraux soumise à un problème de minimisation du risque appelé *modèle à données latentes*. Soit Z un sous-ensemble de $\mathbb{R}^m$, μ une mesure finie σ sur le Borel σ-algebra $\mathcal{Z} = \mathcal{B}(\mathsf{Z})$ et $\{f(z, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ soit une famille de fonctions Borel positives μ-integrable sur Z. Set $z \in \mathsf{Z}$. Soit, pour tout $\boldsymbol{\theta} \in \Theta$:

$$\begin{aligned} g(y;\boldsymbol{\theta}) &\triangleq \int_{\mathsf{Z}} f(z, y; \boldsymbol{\theta}) \mu(\mathrm{d}z) \;, \\ p(z|y;\boldsymbol{\theta}) &\triangleq \begin{cases} \frac{f(z,y;\boldsymbol{\theta})}{g(y;\boldsymbol{\theta})} & \text{if } g(y;\boldsymbol{\theta}) \neq 0 \\ 0 & \text{sinon} \end{cases} \end{aligned} \tag{2.3.1}$$

Notez que $p(z|y;\boldsymbol{\theta})$ définit une fonction de densité de probabilité par rapport à μ et $\mathcal{P} = \{p(z|y;\boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta; (y, z) \in \mathsf{Y} \times \mathsf{Z}\}$ a family of probability density. Nous désignons par $\{\mathbb{P}_{\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ la famille de mesures de probabilité associée. Naturellement, la fonction de

perte $\mathcal{L}(\boldsymbol{\theta})$ est définie pour tous les $\boldsymbol{\theta} \in \Theta$ comme suit :

$$\mathcal{L}(\boldsymbol{\theta}) := \log g(y; \boldsymbol{\theta}) \ . \tag{2.3.2}$$

Remark 2.1 *Un exemple est le problème de données incomplètes. Dans ce cadre,*

- $f(z, y; \boldsymbol{\theta})$ *est la probabilité des données complètes qui est la probabilité des données observées* y *augmentée des données manquantes* z*.*
- $g(y; \boldsymbol{\theta})$ *est la probabilité de données incomplètes qui est la probabilité des données observées* y*.*
- $p(z|y; \boldsymbol{\theta})$ *est la distribution conditionnelle des données manquantes* z *sachant les données observées* y*.*

Remark 2.2 *Pour les modèles à effets mixtes, les variables latentes* z *sont les effets aléatoires et l'identification de la structure latentes correspond principalement à la variabilité inter-individuelle entre les individus du jeu de données. Ce cadre d'étude est présenté Section 1.4 et étudié Chapitre 6.*

Remark 2.3 *Pour les modèles de mélange, les variables latentes correspondent aux labels inconnus du mélange en prenant des valeurs dans un ensemble fini discret. Ce cadre d'analyse est étudié Chapitre 4, Chapitre 5 et Chapitre 7.*

Remark 2.4 *Dans cette thèse, nous nous intéressons à une approche empirique du problème de l'estimation du maximum de vraisemblance. Soit* n *un entier. Nous considérons* n *vecteurs d'observations indépendant et non nécessairement distribué de façon identique* $(y_i \in \mathsf{Y}, i \in [\![1, n]\!])$ *où* Y *est un sous-ensemble de* $\mathbb{R}^{l_i}$ *et données latentes* $(z_i \mathsf{Z}, i \in [\![1, n]\!])$*. Pour tout* $\boldsymbol{\theta} \in \Theta$*,*

$$\begin{aligned} f(z, y; \boldsymbol{\theta}) &= \prod_{i=1}^{n} f(z_i, y_i; \boldsymbol{\theta}) \ , \\ g(y; \boldsymbol{\theta}) &= \prod_{i=1}^{n} g(y_i; \boldsymbol{\theta}) \ , \\ p(z|y; \boldsymbol{\theta}) &= \prod_{i=1}^{n} p(z_i|y_i; \boldsymbol{\theta}) \ . \end{aligned} \tag{2.3.3}$$

Ainsi, la fonction objective (2.3.2) s'écrit :

$$\mathcal{L}(\boldsymbol{\theta}) := \sum_{i=1}^{n} \log g(y_i; \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \ . \tag{2.3.4}$$

Notez que pour éviter les singularités et les dégénérescences du Maximum de Vraisemblance (MV) telles que mises en évidence dans [Fraley and Raftery, 2007], on peut régulariser

la fonction objective par une distribution a priori sur les paramètres du modèle, voir Chapitre 4 pour un exemple illustratif.

2.3.2 L'algorithme EM

Une classe populaire d'algorithmes d'inférence ayant pour but minimiser (2.3.2) est la classe d'algorithmes du type Expectation-Maximization (EM) développé dans le travail pionnier de Dempster et al. [1977]. L'EM est une procédure itérative qui minimise la fonction $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta})$ lorsque sa minimisation directe est difficile. Indiquez par $\boldsymbol{\theta}^{(k-1)}$ le paramètre connu à l'itération k, alors la k-ième étape de l'algorithme EM pourrait être décomposée en deux étapes. L'étape E consiste à calculer la fonction de substitution définie pour tous les $\boldsymbol{\theta} \in \Theta$ comme :

$$Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)}) \triangleq \int_{\mathsf{Z}} p(z|y; \boldsymbol{\theta}^{(k-1)}) \log f(z, y; \boldsymbol{\theta}) \mu(\mathrm{d}z) \; . \tag{2.3.5}$$

Dans l'étape M, la valeur de $\boldsymbol{\theta}$ minimisant $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ est calculée et définie comme la nouvelle estimation de paramètre $\boldsymbol{\theta}^{(k)}$. Ces deux étapes sont répétées jusqu'à la convergence. L'essence de l'algorithme EM est que la diminution de $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ force une diminution de la fonction $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta})$, voir [McLachlan and Krishnan, 2007] et les références qui y figurent.

Remark 2.5 *En utilisant la concavité de la fonction logarithmique et l'inégalité de Jensen, nous pouvons montrer que $Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$ est une fonction de substitution majorante de l'objectif $\mathcal{L}(\boldsymbol{\theta})$ au point $\boldsymbol{\theta}^{(k-1)}$. Ce schéma s'inscrit bien dans le principe MM introduit dans la Section 1.2.1 et est exploité dans [Gunawardana and Byrne, 2005]. Le Chapitre 5 développe cette remarque et présente une analyse globale d'une variante incrémentale de l'EM, introduite par Neal and Hinton [1998].*

Remark 2.6 *Une hypothèse courante concernant l'applicabilité directe de l'EM aux modèles à données latentes (voir, en particulier, la discussion dans l'article [Dempster et al., 1977]) est de considérer que le modèle complet appartient à la famille exponentielle courbe, à savoir, pour tout $\boldsymbol{\theta} \in \Theta$:*

$$\log f(z, y, \boldsymbol{\theta}) = -\psi(\boldsymbol{\theta}) + \langle \tilde{S}(z, y), \phi(\boldsymbol{\theta}) \rangle. \tag{2.3.6}$$

où $\psi : \boldsymbol{\theta} \mapsto \mathbb{R}$ et $\phi : \boldsymbol{\theta} \mapsto \mathbb{R}$ sont des fonctions deux fois continument différentiables en $\boldsymbol{\theta}$ et $\tilde{S} : \mathsf{Z} \mapsto \mathsf{S}$ est une statistique prenant ses valeurs dans un sous-ensemble convexe S de $\mathbb{R}$. Ensuite, les deux étapes de l'EM formulent en termes de statistiques suffisantes. En particulier, l'étape M jouit d'une fonction d'expression fermée de ces statistiques. Notez que cette hypothèse n'est pas restrictive car de nombreux modèles d'intérêt pour l'apprentissage machine la satisfont.

Travaux antérieurs La méthode EM a fait l'objet d'un intérêt considérable depuis son introduction dans [Dempster et al., 1977]. La plupart des travaux traitant de la convergence des méthodes de type EM considèrent les comportements *asymptotique* et/ou *local* pour éviter toute hypothèse de non-convexité. La convergence globale vers un point stationnaire (soit un minimum local, soit un point de selle) de la méthode EM a été établie par Wu et al. [1983] comme une extension des travaux antérieurs développés dans Dempster et al. [1977]. La convergence globale est une conséquence directe de la monotonie de la méthode EM, c'est-à-dire que la fonction objective ne décroît jamais. Localement et sous certaines conditions de régularité, un taux de convergence linéaire vers un point stationnaire a été étudié dans [McLachlan and Krishnan, 2007, Chapitres 3 et 4]. En ce qui concerne les variantes incrémentales, la convergence de la méthode iEM a d'abord été abordée par Gunawardana and Byrne [2005] en exploitant l'interprétation de la méthode comme une procédure de minimisation alternée dans le cadre de l'Information Géométrique développé dans [Csiszár and Tusnády, 1984]. Plus récemment, l'accent mis sur la convergence locale mais non asymptotique des méthodes EM a été étudié dans plusieurs travaux. Ces résultats exigent généralement d'initialiser l'algorithme au voisinage d'un point stationnaire isolé et que la fonction de log-vraisemblance (négative) soit fortement convexe localement. Ces conditions sont difficiles à vérifier en général ou n'ont été dérivées que pour des modèles spécifiques ; voir par exemple [Balakrishnan et al., 2017, Wang et al., 2015a, Xu et al., 2016a] et les références qui y figurent. La convergence locale d'une méthode EM à variance réduite, appelée sEM-VR a été étudiée dans [Chen et al., 2018, Theorem 1] mais sous une condition de stabilité globale.

Il est donc important d'améliorer et d'analyser les variantes de l'EM afin de relever les deux défis mentionnés au tout début de cette introduction, à savoir le nombre croissant de données et la non-convexité de la fonction objective (voir Chapitre 5 du manuscrit).

2.3.3 L'algorithme SAEM

Dans de nombreuses situations, l'étape de calcul de l'espérance dans l'algorithme EM (2.3.5) peut être numériquement compliquée ou même intractable. Pour résoudre ce problème, Wei and Tanner [1990a] propose de remplacer le terme d'expérance par une intégration de Monte Carlo, conduisant à ce que l'on appelle l'algorithme Monte Carlo EM (MCEM). Une autre option, développée dans [Delyon et al., 1999a], est l'approximation stochastique de l'EM (SAEM) qui s'écrit comme suit :

1. **Etape de Simulation:** Simulez les variables latentes $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ à partir de sa distribution a posteriori $p(z|y;\boldsymbol{\theta}^{(k-1)})$.

2. **Etape d'Approximation Stochastique:** mettre à jour l'approximation, notée $Q_k(\boldsymbol{\theta})$, de l'espérance conditionnelle (2.3.5) :

$$Q_k(\boldsymbol{\theta}) = Q_{k-1}(\boldsymbol{\theta}) + \gamma_k \left(M_{(k)}^{-1} \sum_{m=1}^{M_{(k)}} \log f(y, z_m^{(k)}; \boldsymbol{\theta}) - Q_{k-1}(\boldsymbol{\theta}) \right) , \tag{2.3.7}$$

où $\{\gamma_k\}_{k>0}$ est une séquence de pas décroissants avec $\gamma_1 = 1$.

3. **Etape de Maximisation:**

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}\in\Theta} Q_k(\boldsymbol{\theta}) . \tag{2.3.8}$$

Pendant la phase d'approximation stochastique, la distribution conditionnelle des paramètres est obtenue puisqu'il s'agit de la distribution dans laquelle les variables latentes z sont imputées pour obtenir un ensemble de données complet à partir duquel est dérivé le log-vraisemblance conditionnelle (voir [Kuhn and Lavielle, 2004]).

Dans l'étape de simulation, comme la relation entre les données observées et les données latentes peut être non linéaire, l'échantillonnage de la distribution postérieure est difficile et nécessite souvent l'utilisation d'un algorithme d'inférence. Kuhn and Lavielle [2004] a démontré une convergence quasi certaine de la séquence de paramètres obtenue par cet algorithme couplée à une procédure MCMC pendant l'étape de simulation. En effet, ici, $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ est un ensemble d'échantillons de Monte Carlo. Dans des scénarios simples, les échantillons $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ sont conditionnellement indépendants et distribués de manière identique selon la distribution $p(z|y, \boldsymbol{\theta}^{(k-1)})$. Néanmoins, dans la plupart des cas, l'échantillonnage exact à partir de cette distribution n'est pas une option et l'ensemble de Monte Carlo est échantillonné par l'algorithme MCMC.

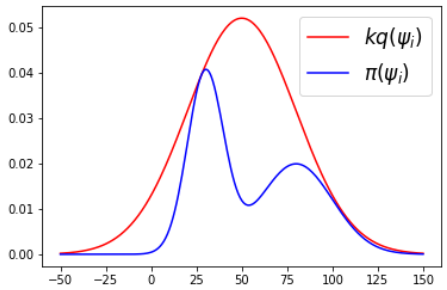

Figure 2.2 – Algorithme MH: représentation d'une distribution de proposition et d'une distribution cible en dimension 1.

Les algorithmes MCMC sont une classe de méthodes permettant d'échantillonner à partir d'une distribution complexe sur (éventuellement) un grand espace dimensionnel. Une classe importante d'échantillonneurs, appelée algorithme Metropolis-Hastings (MH), prélève itérativement des échantillons à partir d'une distribution de proposition q avec la

distribution de l'échantillon nouvellement prélevé seulement selon l'échantillon courant. Avec une certaine probabilité, l'échantillon est soit accepté comme le nouvel état de la chaîne, soit rejeté. Il est bien connu, voir [Mengersen and Tweedie, 1996, Roberts and Rosenthal, 2011] que l'échantillonneur indépendant est géométriquement ergodique si et seulement si, pour un ε donné, $\inf_{z\in\mathbb{R}^p} q(z)/\pi(z) \geq \varepsilon > 0$ où $\pi(z)$ est la distribution cible. Plus généralement, il est montré dans [Roberts and Rosenthal, 2011] que le taux de mélange dans la variation totale dépend de l'espérance du taux d'acceptation dans la distribution de la proposition qui est également directement lié au rapport entre la proposition et la cible. Cette observation suggère naturellement de trouver une proposition qui se rapproche de l'objectif. La figure 2.2 illustre cette remarque lorsque la proposition est une simple distribution gaussienne. à partir de cette figure, on peut reconnaêtre que l'efficacité de l'échantillonneur sera influencée par le niveau de similarité (*eg.* ils appartiennent à la même famille de distributions) entre les deux distributions de l'intérêt. Le Chapitre 6 développe cette remarque en présentant une proposition MH efficace pour la tâche d'échantillonnage.

Dans l'étape d'approximation stochastique, la séquence de pas positifs et décroissants $\{\gamma_k\}_{k>0}$ contrôle la convergence de l'algorithme. En pratique, γ_k est égal à 1 lors des premières itérations K_1 pour permettre à l'algorithme d'explorer l'espace de paramètres sans mémoire et de converger rapidement vers un voisinage du MV. L'approximation stochastique est effectuée lors des itérations finales K_2 où $\gamma_k = 1/k^a$ avec en général $a = 0.7$, assurant la convergence presque certaine de l'estimateur.

Travaux antérieurs Il a été démontré que l'algorithme SAEM converge théoriquement vers un maximum de vraisemblance des observations dans des conditions générales [Delyon et al., 1999a]. Comme déjà mentionné, ce résultat a été étendu par Kuhn and Lavielle [2004], pour inclure une procédure d'échantillonnage MCMC dans la phase de simulation. Les travaux récents de Allassonnière and Chevallier [2019], présentent une nouvelle classe d'algorithmes où l'étape de simulation est effectuée en utilisant une version tempérée de la distribution postérieure et est motivée par les problèmes des points de selle.

2.4 Modèles à Effets Mixtes et Approche de Population

2.4.1 Pourquoi Les Modèles à Effets Mixtes Sont-ils Pertinents?

Les modèles à effets mixtes (MEM), voir [Lavielle, 2014] et les références qui y figurent, sont de plus en plus utilisés en raison de leur souplesse à analyser des données longitudinales multiples selon des profils possiblement non linéaires. Ce sont des méthodes de référence pour décrire la variabilité inter-individuelle au sein d'une population.

Une formulation générale des MEM pour une observation continue y_{ij} peut s'écrire comme

suit :

$$y_{ij} = \mathsf{f}(x_{ij}, \psi_i) + \mathsf{g}(x_{ij}, \psi_i, \xi_i)\, \varepsilon_{ij} \quad \text{avec} \quad \varepsilon_{ij} \sim \mathcal{N}(0, \sigma^2)\ , \tag{2.4.1}$$

où la quantité y_{ij} désigne la j-ième observation pour le i-ième individu. x_{ij} est un vecteur de variables explicatives (il peut s'agir du temps ou de certaines variables telles que le temps ou la dose d'un médicament injecté), f est le modèle structurel (éventuellement non linéaire), ψ_i désigne les paramètres individuels. La quantité ε_{ij} est une variable aléatoire supposée être normalement distribuée et σ désigne le paramètre de variance entrant dans la fonction g, qui exprime l'écart-type de l'erreur de mesure et est généralement soit constante (variance homoscédastique) ou une fonction de f.

Nous considérons ici un modèle en deux étapes, comme dans [Davidian, 2017], qui fournit une courbe de population typique, également connue sous le nom de *modèle structurel.* (voir la fonction f dans (2.4.1)), et modélise les paramètres individuels, dénotés ψ_i et considérés comme des variables aléatoires qui fluctuent autour d'un paramètre population ψ_{pop}. Ce dernier modèle probabiliste des paramètres individuels montre la structure de variabilité inter-individuelle qui régit les phénomènes statistiques. Formellement, il se lit comme suit :

$$\begin{cases} \psi_i = \mathsf{h}(\psi_{\text{pop}}, \eta_i) \\ \eta_i \sim \mathcal{N}(0, \Omega) \end{cases} \tag{2.4.2}$$

Nous notons que les paramètres individuels ψ_i sont liés par une fonction h à ψ_{pop}, le vecteur p-dimensionnel contenant les effets fixes, et η_i, le vecteur q-dimensionnel contenant les effets aléatoires. De plus, ε_{ij} est une variable aléatoire supposée être normalement distribuée et σ indique la variance entrant dans la fonction g, qui exprime l'écart-type de l'erreur de mesure et est généralement soit constante (variance homoscédastique) ou une fonction du modèle structurel f. Les effets aléatoires η_i et les erreurs résiduelles ε_{ij} sont supposés être indépendants pour différents sujets et indépendants les uns des autres pour le même sujet.

L'objectif ici est d'etimer le vecteur de paramètres $\boldsymbol{\theta} = (\psi_{\text{pop}}, \Omega, \sigma^2)$ par maximum de vraisemblance. Dans les modèles à effets mixtes, la probabilité associée à (2.4.1) et (2.4.2) est intractable car les probabilités individuelles doivent intégrer les paramètres inconnus ψ_i selon leur distribution. Les paramètres individuels estimés, appelés estimations empiriques de Bayes (EBE), peuvent être définis comme le mode ou la médiane de la distribution conditionnelle.

Pour les MEM, l'espérance calculée en (2.3.5) est intractable en raison de la non-linéarité possible du modèle structurel. Nous utilisons donc l'algorithme SAEM présenté ci-dessus où les variables latentes, qui sont simulées à chaque itération, correspondent aux paramètres individuels ψ_i.

2.4.2 Applications en Pharmacocinétique

Dans des domaines tels que l'économie, la sociologie, la génomique ou la pharmacocinétique (PK), on mesure les observations de plusieurs individus d'une même population. Considérons les observations Figure 2.3.

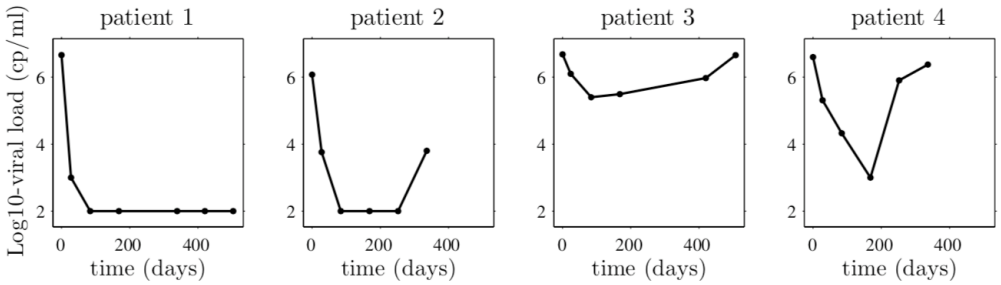

Figure 2.3 – Charge virale pour 4 patients atteints d'hepatitis C (tiré de [Lavielle, 2014]).

Ces mesures sont des charges virales pour quatre patients différents atteints d'hépatite C (VHC) après un traitement qui a commencé au moment $t = 0$. Nous notons que dans de nombreux cas, comme celui exposé ici, les données sont des mesures répétées (pas nécessairement en fonction du temps) d'une grandeur de sortie, c'est-à-dire qu'elles sont *longitudinales*. De plus, chaque patient ne réagit pas de la même façon au traitement. Par exemple, le patient 1 voit sa charge virale diminuer après le traitement alors que le patient 3 n'a presque aucune réponse. Suite à ces deux remarques, la meilleure façon de faire face à la modélisation statistique de la dynamique du VHC est de construire un modèle qui décrit comment la charge virale évolue dans le temps et un modèle statistique qui explique la différence entre les patients. La modélisation des effets mixtes est un cadre naturel pour une telle analyse et a été développée en profondeur dans [Snoeck et al., 2010] pour la modélisation dynamique du VHC.

L'approche de population devient alors pertinente car elle combine (possiblement) de pauvres informations individuelles pour construire un modèle de population complet et riche.

Les MEM et l'algorithme SAEM sont largement utilisés pour traiter ce type d'éxample à travers leur implémentation dans des outils logiciels tels que Monolix, NONMEM, le package R SAEMIX [Comets et al., 2017] et la fonction Matlab nlmefitsa. Une partie de notre travail dans cette thèse repose sur le package SAEMIX (R [R Development Core Team, 2008]), voir Chapitre 6-8.

Part I

NON-CONVEX RISK MINIMIZATION

Chapter 3

Incremental Method for Non-smooth Non-convex Optimization

Abstract: *Many constrained, non-convex optimization problems can be tackled using the Majorization-Minimization (MM) method which alternates between constructing a surrogate function which upper bounds the objective function, and then minimizing this surrogate. For problems which minimize a finite sum of functions, a stochastic version of the MM method selects a batch of functions at random at each iteration and optimizes the accumulated surrogate. However, in many cases of interest such as variational inference for latent variable models, the surrogate functions are expressed as an expectation. In this contribution, we propose a doubly stochastic MM method based on Monte Carlo approximation of these stochastic surrogates. We establish asymptotic and non-asymptotic convergence of our scheme in a constrained, non-convex, non-smooth optimization setting. We apply our new framework for inference of logistic regression model with missing covariates and for variational inference of autoencoder on the MNIST dataset. This chapter corresponds to the article [Karimi et al., 2019b].*

Contents

3.1 Introduction

We consider the *constrained* minimization problem of a finite sum of functions:

$$\min_{\boldsymbol{\theta} \in \Theta} \; \mathcal{L}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \; , \tag{3.1.1}$$

where Θ is a convex, compact, and closed subset of $\mathbb{R}^p$, and for any $i \in [\![1, n]\!]$, the function $\mathcal{L}_i : \mathbb{R}^p \to \mathbb{R}$ is bounded from below and is (possibly) non-convex and non-smooth.

To tackle the optimization problem (3.1.1), a popular approach is to apply the majorization-minimization (MM) method which iteratively minimizes a majorizing surrogate function. A large number of existing procedures fall into this general framework, for instance gradient-based or proximal methods or the Expectation-Maximization (EM) algorithm [McLachlan and Krishnan, 2008] and some variational Bayes inference techniques [Jordan et al., 1999]; see for example [Razaviyayn et al., 2013] and [Lange, 2016] and the references therein. When the number of terms n in (3.1.1) is large, the vanilla MM method may be intractable because it requires to construct a surrogate function for all the n terms $\mathcal{L}_i$ at each iteration. Here, a remedy is to apply the Minimization by Incremental Surrogate Optimization (MISO) method proposed by Mairal [2015b], where the surrogate functions are updated incrementally. The MISO method can be interpreted as a combination of MM and ideas which have emerged for variance reduction in stochastic gradient methods [Schmidt et al., 2017].

The success of the MISO method rests upon the efficient minimization of surrogates such as convex functions, see [Mairal, 2015b, Section 2.3]. In many applications of interest, the natural surrogate functions are intractable, yet they are defined as expectation of tractable functions. This for example the case for inference in latent variable models. Another application is variational inference, [Ghahramani, 2015], in which the goal is to approximate the posterior distribution of parameters given the observations; see for example [Blundell et al., 2015, Li and Gal, 2017, Neal, 2012, Polson et al., 2017, Rezende et al., 2014].

This paper fills the gap in the literature by proposing a new method called *Minimization by Incremental Stochastic Surrogate Optimization (MISSO)* which is designed for the finite sum optimization with a finite-time convergence guarantee. Our contributions can be summarized as follows.

- We propose a unifying framework of analysis for incremental stochastic surrogate optimization when the surrogates are defined by expectations of tractable functions. The proposed MISSO method is built on the Monte Carlo integration of the intractable surrogate function, *i.e.,* a doubly stochastic surrogate optimization scheme. In addition, we present an incremental variational inference and Monte-Carlo EM methods as two special cases of this framework.
- We establish both asymptotic and non-asymptotic convergence for the MISSO method. In particular, the MISSO method converges almost surely to a stationary point and in $\mathcal{O}(n/\epsilon)$ iterations to an ϵ-stationary point.

In Section 3.2, we review the techniques for incremental minimization of finite sum functions based on the MM principle; specifically, we review the MISO method as introduced in [Mairal, 2015b], and present a class of surrogate functions expressed as an expectation over a latent space. The MISSO method is then introduced for the latter class of surrogate functions. In Section 4.2.1, we provide the asymptotic and non-asymptotic convergence analysis for the MISSO method. Finally, Section 3.4 presents numerical applications to illustrate our findings including parameter inference for logistic regression with missing covariates and variational inference for Bayesian neural network.

Notations We denote $[\![1, n]\!] = \{1, \ldots, n\}$. Unless otherwise specified, $\|\cdot\|$ denotes the standard Euclidean norm and $\langle \cdot \,|\, \cdot \rangle$ is the inner product in Euclidean space. For any function $f : \Theta \to \mathbb{R}$, $f'(\boldsymbol{\theta}, \boldsymbol{d})$ is the directional derivative of f at $\boldsymbol{\theta}$ along the direction $\boldsymbol{d}$, *i.e.,*

$$f'(\boldsymbol{\theta}, \boldsymbol{d}) := \lim_{t \to 0^+} \frac{f(\boldsymbol{\theta} + t\boldsymbol{d}) - f(\boldsymbol{\theta})}{t} \,. \tag{3.1.2}$$

The directional derivative is assumed to exist for the functions introduced throughout this paper.

3.2 Incremental Minimization of Finite Sum Non-convex Functions

The objective function in (3.1.1) is composed of a finite sum of possibly non-smooth and non-convex functions. A popular approach here is to apply the MM method. The MM method tackles (3.1.1) through alternating between two steps — (i) minimizing a *surrogate* function which upper bounds the original objective function; and (ii) updating the surrogate function to tighten the upper bound.

As mentioned in the Introduction, the MISO method proposed by Mairal [2015b] is developed as an iterative scheme that only updates the surrogate functions *partially* at each

iteration. Formally, for any $i \in [\![1, n]\!]$, we consider a surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ which satisfies

S3.1 *For all $i \in [\![1, n]\!]$ and $\overline{\boldsymbol{\theta}} \in \Theta$, the function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ is convex* w.r.t. *$\boldsymbol{\theta}$, and it holds*

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta}), \ \forall \ \boldsymbol{\theta} \in \Theta \ , \tag{3.2.1}$$

where the equality holds when $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$.

S3.2 *For any $\overline{\boldsymbol{\theta}}_i \in \Theta$, $i \in [\![1, n]\!]$ and some $\epsilon > 0$, the difference function $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}_i) - \mathcal{L}(\boldsymbol{\theta})$ is defined for all $\boldsymbol{\theta} \in \Theta_\epsilon$ and differentiable for all $\boldsymbol{\theta} \in \Theta$, where $\Theta_\epsilon = \{\boldsymbol{\theta} \in \mathbb{R}^d, \inf_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \epsilon\}$ is an ϵ-neighborhood set of Θ. Moreover, for some constant L, the gradient satisfies*

$$\|\nabla \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)\|^2 \leq 2L \, \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n), \ \forall \ \boldsymbol{\theta} \in \Theta \ . \tag{3.2.2}$$

S3.1 is a common condition used for surrogate optimization, see [Mairal, 2015b, Section 2.3]. Meanwhile, S3.2 can be satisfied when the difference function $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)$ is L-smooth for all $\boldsymbol{\theta} \in \mathbb{R}^d$, where the condition can be implied through applying [Razaviyayn et al., 2013, Proposition 1].

Algorithm 3.1 MISO method [Mairal, 2015b]

1: **Input:** initialization $\boldsymbol{\theta}^{(0)}$.
2: Initialize the surrogate function as $\mathcal{A}_i^0(\boldsymbol{\theta}) := \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}), \ i \in [\![1, n]\!]$.
3: **for** $k = 0, 1, \ldots$ **do**
4: Pick i_k uniformly from $[\![1, n]\!]$.
5: Update $\mathcal{A}_i^{k+1}(\boldsymbol{\theta})$ as:

$$\mathcal{A}_i^{k+1}(\boldsymbol{\theta}) = \begin{cases} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}), & \text{if } i = i_k \\ \mathcal{A}_i^k(\boldsymbol{\theta}), & \text{otherwise.} \end{cases}$$

6: Set $\boldsymbol{\theta}^{(k+1)} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \frac{1}{n} \sum_{i=1}^n \mathcal{A}_i^{k+1}(\boldsymbol{\theta})$.
7: **end for**

The inequality (4.2.11) implies $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta}) > -\infty$ for any $\boldsymbol{\theta} \in \Theta$. The MISO method is an incremental version of the MM method, as summarized by 3.1. As seen in the pseudo code, the MISO method maintains an iteratively updated set of surrogate upper-bound functions $\{\mathcal{A}_i^k(\boldsymbol{\theta})\}_{i=1}^n$ and updates the iterate through minimizing the average of the surrogate functions.

Particularly, only one out of the n surrogate functions is updated at each iteration [cf. Line 5] and the sum function $\frac{1}{n} \sum_{i=1}^n \mathcal{A}_i^{k+1}(\boldsymbol{\theta})$ is designed to be 'easy to optimize', for example, it can be a sum of quadratic functions. As such, the MISO method is suitable for large-scale optimization as the computation cost per iteration is independent of n.

Moreover, under S3.1, S3.2, it was shown that the MISO method converges almost surely to a stationary point of (3.1.1) [Mairal, 2015b, Proposition 3.1].

We now consider the case when the surrogate functions $\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}})$ are intractable. Let Z be a measurable set, $p_i : \mathsf{Z} \times \Theta \to \mathbb{R}_+$ be a pdf, $r_i : \Theta \times \Theta \times \mathsf{Z} \to \mathbb{R}$ be a measurable function and μ_i be a σ-finite measure, we consider surrogate functions which satisfy S3.1, S3.2 that can be expressed as an expectation:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}) := \int_{\mathsf{Z}} r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z_i) p_i(z_i;\overline{\boldsymbol{\theta}}) \mu_i(dz_i) \quad \forall\ (\boldsymbol{\theta},\overline{\boldsymbol{\theta}}) \in \Theta \times \Theta\ . \tag{3.2.3}$$

Plugging (3.2.3) into the MISO method is not feasible since the update step in Step 6 involves a minimization of an expectation. Several motivating examples of (3.1.1) are given in Section 3.2.

We propose the *Minimization by Incremental Stochastic Surrogate Optimization* (MISSO) method which replaces the expectation in (3.2.3) by *Monte Carlo* integration and then optimizes (3.1.1) incrementally. Denote by $M \in \mathbb{N}$ the Monte Carlo batch size and let $z_{i,m} \in \mathsf{Z}$, $m = 1, ..., M$ be a set of samples for all $i \in [\![1, n]\!]$. These samples can be drawn (Case 1) i.i.d. from the distribution $p_i(\cdot;\overline{\boldsymbol{\theta}})$ or (Case 2) from a Markov chain with the stationary distribution $p_i(\cdot;\overline{\boldsymbol{\theta}})$; see Section 4.2.1 for illustrations. To this end, we define

$$\widetilde{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, \{z_{i,m}\}_{m=1}^M) := \frac{1}{M} \sum_{m=1}^{M} r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z_{i,m}) \tag{3.2.4}$$

and we summarize the proposed MISSO method in 3.2. As seen, the procedure is similar to the MISO method but it involves two types of randomness. The first randomness comes from the selection of i_k in Line 5. The second randomness is that a set of Monte-Carlo approximated functions $\widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta})$ is used in lieu of $\mathcal{A}_i^k(\boldsymbol{\theta})$ when optimizing for the next iterate $\boldsymbol{\theta}^{(k)}$. We now discuss two applications of the MISSO method.

Example 1: Maximum Likelihood Estimation for Latent Variable Model Latent variable models [Bishop, 2006] are constructed by introducing unobserved (latent) variables which help explain the observed data. We consider n independent observations $((y_i, z_i), i \in [\![n]\!])$, that can be non identically distributed, where y_i is observed and z_i is latent. In this incomplete data framework, define $\{f_i(z_i, y_i, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ to be the complete data likelihood models, *i.e.,* joint likelihood of the observations and latent variables. Let

$$g_i(y_i, \boldsymbol{\theta}) := \int_{\mathsf{Z}} f_i(z_i, y_i, \boldsymbol{\theta}) \mu_i(\mathrm{d}z_i),\ i \in [\![1, n]\!] \tag{3.2.7}$$

denote the incomplete data likelihood, *i.e.,* the marginal likelihood of the observations. For ease of notations, the dependence on the observations is made implicit. The maximum likelihood (ML) estimation problem takes $\mathcal{L}_i(\boldsymbol{\theta})$ to be the ith negated incomplete data

Algorithm 3.2 MISSO method

Input: initialization $\boldsymbol{\theta}^{(0)}$; a sequence of non-negative numbers $\{M_{(k)}\}_{k=0}^{\infty}$.
For all $i \in [\![1, n]\!]$, draw $M_{(0)}$ Monte-Carlo samples with the stationary distribution $p_i(\cdot; \boldsymbol{\theta}^{(0)})$.
Initialize the surrogate function as

$$\widetilde{\mathcal{A}}_i^0(\boldsymbol{\theta}) := \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(k)}}), \ i \in [\![1, n]\!] \ . \tag{3.2.5}$$

for $k = 0, 1, \ldots$ **do**
  Pick a function index i_k uniformly on $[\![1, n]\!]$.
  Draw $M_{(k)}$ Monte-Carlo samples with the stationary distribution $p_i(\cdot; \boldsymbol{\theta}^{(k)})$.
  Update the individual surrogate functions recursively as:

$$\widetilde{\mathcal{A}}_i^{k+1}(\boldsymbol{\theta}) = \begin{cases} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}), & \text{if } i = i_k \\ \widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta}), & \text{otherwise.} \end{cases} \tag{3.2.6}$$

  Set $\boldsymbol{\theta}^{(k+1)} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \widetilde{\mathcal{A}}_i^{k+1}(\boldsymbol{\theta})$.
end for

log-likelihood $\mathcal{L}_i(\boldsymbol{\theta}) := -\log g_i(y_i, \boldsymbol{\theta})$.

Assume without loss of generality that $g_i(y_i, \boldsymbol{\theta}) \neq 0$ for all $\boldsymbol{\theta} \in \Theta$, we define by $p_i(z_i|y_i, \boldsymbol{\theta}) := f_i(z_i, y_i, \boldsymbol{\theta})/g_i(y_i, \boldsymbol{\theta})$ the conditional distribution of the latent variable z_i given the observation y_i. A surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ satisfying S3.1 can be obtained through writing $f_i(z_i, y_i, \boldsymbol{\theta}) = \frac{f_i(z_i, y_i, \boldsymbol{\theta})}{p_i(z_i|y_i, \overline{\boldsymbol{\theta}})} p_i(z_i|y_i, \overline{\boldsymbol{\theta}})$ and applying the Jensen inequality:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) = \int_{\mathsf{Z}} \underbrace{\log\left(p_i(z_i, \overline{\boldsymbol{\theta}})/f_i(z_i, y_i, \boldsymbol{\theta})\right)}_{= r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i)} p_i(z_i|y_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i) \ , \tag{3.2.8}$$

We note that S3.2 can also be verified for common distribution models. We can apply the MISSO method following the above specification of $r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i), p_i(z_i|y_i, \overline{\boldsymbol{\theta}})$.

Example 2: Variational Inference Let $((x_i, y_i), i \in [\![1, n]\!])$ be i.i.d. input-output pairs and $w \in \mathsf{W} \subseteq \mathbb{R}^d$ be a latent variable. When conditioned on the input $x = (x_i, i \in [\![1, n]\!])$, the joint distribution of $y = (y_i, i \in [\![1, n]\!])$ and w is given by:

$$p(y, w|x) = \pi(w) \prod_{i=1}^n p(y_i|x_i, w) \ . \tag{3.2.9}$$

Our goal is to compute the posterior distribution $p(w|y, x)$. In most cases, the posterior distribution $p(w|y, x)$ is intractable and is approximated using a family of parametric distributions, $\{q(w, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$. The variational inference (VI) problem [Blei et al., 2017a] boils down to minimizing the KL divergence between $q(w, \boldsymbol{\theta})$ and the posterior distribution

$p(w|y,x)$, as follows:

$$\min_{\boldsymbol{\theta}\in\Theta} \mathcal{L}(\boldsymbol{\theta}) := \mathrm{KL}\left(q(w;\boldsymbol{\theta}) \,||\, p(w|y,x)\right) := \mathbb{E}_{q(w;\boldsymbol{\theta})}\left[\log\left(q(w;\boldsymbol{\theta})/p(w|y,x)\right)\right] . \tag{3.2.10}$$

Using (3.2.9), we decompose $\mathcal{L}(\boldsymbol{\theta}) = n^{-1}\sum_{i=1}^{n}\mathcal{L}_i(\boldsymbol{\theta}) + \text{const.}$ where:

$$\mathcal{L}_i(\boldsymbol{\theta}) := -\mathbb{E}_{q(w;\boldsymbol{\theta})}\left[\log p(y_i|x_i,w)\right] + \frac{1}{n}\mathbb{E}_{q(w;\boldsymbol{\theta})}\left[\log q(w;\boldsymbol{\theta})/\pi(w)\right] = r_i(\boldsymbol{\theta}) + d(\boldsymbol{\theta}) . \tag{3.2.11}$$

Directly optimizing the finite sum objective function in (3.2.10) can be difficult. First, with $n \gg 1$, evaluating the objective function $\mathcal{L}(\boldsymbol{\theta})$ requires a full pass over the entire dataset. Second, for some complex models, the expectations in (3.2.11) can be intractable even if we assume a simple parametric model for $q(w;\boldsymbol{\theta})$. Assume that $\mathcal{L}_i$ is L-smooth, *i.e.,* $\mathcal{L}_i$ is differentiable on Θ and its gradient $\nabla\mathcal{L}_i$ is L-Lipschitz. We apply the MISSO method with a quadratic surrogate function defined as:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}) := \mathcal{L}_i(\overline{\boldsymbol{\theta}}) + \left\langle \nabla_{\boldsymbol{\theta}}\mathcal{L}_i(\overline{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{\mathrm{L}}{2}\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 . \tag{3.2.12}$$

It is easily checked that $\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}})$ satisfies S3.1, S3.2. To compute the gradient $\nabla\mathcal{L}_i(\overline{\boldsymbol{\theta}})$, we apply the re-parametrization technique suggested in [Blundell et al., 2015, Kingma and Welling, 2014, Paisley et al., 2012]. Let $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$ be a differentiable function *w.r.t.* $\boldsymbol{\theta} \in \Theta$ which is designed such that the law of $w = t(z,\overline{\boldsymbol{\theta}})$, where $z \sim \mathcal{N}_d(0,\mathbf{I})$, is $q(\cdot,\overline{\boldsymbol{\theta}})$. By [Blundell et al., 2015, Proposition 1], the gradient of $-r_i(\cdot)$ in (3.2.11) is:

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q(w;\overline{\boldsymbol{\theta}})}\left[\log p(y_i|x_i,w)\right] = \mathbb{E}_{z\sim\mathcal{N}_d(0,\mathbf{I})}\left[\mathrm{J}^t_{\boldsymbol{\theta}}(z,\overline{\boldsymbol{\theta}})\nabla_w \log p(y_i|x_i,w)\big|_{w=t(z,\overline{\boldsymbol{\theta}})}\right] , \tag{3.2.13}$$

where for each $z \in \mathbb{R}^d$, $\mathrm{J}^t_{\boldsymbol{\theta}}(z,\overline{\boldsymbol{\theta}})$ is the Jacobian of the function $t(z,\cdot)$ with respect to $\boldsymbol{\theta}$ evaluated at $\overline{\boldsymbol{\theta}}$. In addition, for most cases, the term $\nabla d(\overline{\boldsymbol{\theta}})$ can be evaluated in closed form.

$$r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}},z) := \left\langle \nabla_{\boldsymbol{\theta}} d(\overline{\boldsymbol{\theta}}) - \mathrm{J}^t_{\boldsymbol{\theta}}(z,\overline{\boldsymbol{\theta}})\nabla_w \log p(y_i|x_i,w)\big|_{w=t(z,\overline{\boldsymbol{\theta}})} \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|^2 .$$

(3.2.14)

Finally, using (3.2.12) and (3.2.14), the surrogate function (3.2.4) is given by $\widetilde{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}},\{z_m\}_{m=1}^M) := M^{-1}\sum_{m=1}^M r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}},z_m)$ where $\{z_m\}_{m=1}^M$ is an i.i.d sample from $\mathcal{N}(0,\mathbf{I})$.

3.3 Convergence Analysis

We provide non-asymptotic convergence bound for the MISSO method and show that it converges asymptotically to a stationary point. Consider the following assumptions.

H3.1 *For all $i \in [\![1, n]\!]$, $\overline{\boldsymbol{\theta}} \in \Theta$, $z_i \in \mathsf{Z}$, the measurable function $r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i)$ is convex in $\boldsymbol{\theta}$ and is lower bounded.*

H3.2 *For the samples $\{z_{i,m}\}_{m=1}^M$, there exists finite constants C_{r} and C_{gr} such that*

$$C_{\mathsf{r}} := \sup_{\overline{\boldsymbol{\theta}} \in \Theta} \sup_{M>0} \frac{1}{\sqrt{M}} \mathbb{E}_{\overline{\boldsymbol{\theta}}} \left[\sup_{\boldsymbol{\theta} \in \Theta} \left| \sum_{m=1}^{M} \left\{ r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_{i,m}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \right\} \right| \right] \tag{3.3.1}$$

$$C_{\mathsf{gr}} := \sup_{\overline{\boldsymbol{\theta}} \in \Theta} \sup_{M>0} \sqrt{M} \mathbb{E}_{\overline{\boldsymbol{\theta}}} \left[\sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{M} \sum_{m=1}^{M} \frac{\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}) - r_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}, z_{i,m})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \right|^2 \right] \tag{3.3.2}$$

for all $i \in [\![1, n]\!]$, and we denoted by $\mathbb{E}_{\overline{\boldsymbol{\theta}}}[\cdot]$ the expectation w.r.t. *a Markov chain $\{z_{i,m}\}_{m=1}^M$ with initial distribution $\xi_i(\cdot; \overline{\boldsymbol{\theta}})$, transition kernel $P_{i,\overline{\boldsymbol{\theta}}}$, and stationary distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$.*

H3.2 essentially requires to control the expectation of the supremum of an empirical process [Boucheron et al., 2013, Shapiro et al., 2009]. In particular, if $M \to \infty$, the surrogate function's value and its directional derivative approximate that of $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ uniformly for all $\boldsymbol{\theta} \in \Theta$. As discussed before, there are two relevant cases here:

Case 1: When the samples $\{z_m\}_{m=1}^M$ used to construct the approximation $\widetilde{\mathcal{L}}_i(\cdot; \cdot,)$ are drawn i.i.d. directly from $p_i(\cdot; \overline{\boldsymbol{\theta}})$ and Θ is bounded, then H3.2 can be implied by the concentration of measure under certain additional regularity conditions.

Case 2: When the samples are generated by an MCMC procedure, H3.2 can be achieved through an maximal inequality for beta-mixing sequences obtained in [Doukhan et al., 1995]. The condition may also be implied by a number of drift and minorization conditions [Meyn and Tweedie, 2012].

Stationarity measure As problem (3.1.1) is a constrained optimization, we consider the following stationarity measure:

$$g(\overline{\boldsymbol{\theta}}) := \inf_{\boldsymbol{\theta} \in \Theta} \frac{\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \quad \text{and} \quad g(\overline{\boldsymbol{\theta}}) = g_+(\overline{\boldsymbol{\theta}}) - g_-(\overline{\boldsymbol{\theta}}) \,, \tag{3.3.3}$$

where $g_+(\overline{\boldsymbol{\theta}}) := \max\{0, g(\overline{\boldsymbol{\theta}})\}$, $g_-(\overline{\boldsymbol{\theta}}) := -\min\{0, g(\overline{\boldsymbol{\theta}})\}$ denote the positive and negative part of $g(\overline{\boldsymbol{\theta}})$, respectively. Note that $\overline{\boldsymbol{\theta}}$ is a stationary point if and only if $g_-(\overline{\boldsymbol{\theta}}) = 0$ [Conn et al., 1993]. Furthermore, suppose that the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k \geq 0}$ has a limit point $\overline{\boldsymbol{\theta}}$ that is a stationary point, then one has $\lim_{k \to \infty} g_-(\boldsymbol{\theta}^{(k)}) = 0$. In this sense, the sequence

$\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ is said to satisfy an *asymptotic stationary point condition*. This is equivalent to [Mairal, 2015b, Definition 2.4].

To explain the condition (3.3.3), observe that if $\overline{\boldsymbol{\theta}} \in \mathrm{int}(\Theta)$, the directional derivative can be replaced by the inner product between the gradient $\nabla\mathcal{L}(\overline{\boldsymbol{\theta}})$ and $\boldsymbol{\theta}-\overline{\boldsymbol{\theta}}$, *i.e.*, $\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta}-\overline{\boldsymbol{\theta}}) = \left\langle \nabla\mathcal{L}(\overline{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta}-\overline{\boldsymbol{\theta}} \right\rangle$. Therefore, from the definition we have $g(\overline{\boldsymbol{\theta}}) = -\|\nabla\mathcal{L}(\overline{\boldsymbol{\theta}})\| = -g_-(\overline{\boldsymbol{\theta}})$. If in addition $g_-(\overline{\boldsymbol{\theta}}) = 0$, then $\overline{\boldsymbol{\theta}}$ is a stationary point to (3.1.1) in the same sense as in unconstrained optimization.

To facilitate our analysis, we define τ_i^k as the iteration index where the ith function is last accessed in the MISSO method prior to iteration k. For example, we have $\tau_{i_k}^{k+1} = k$. We define:

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{3.3.4}$$

We first establish a non-asymptotic convergence rate for the MISSO method:

Theorem 1 *Under S3.1, S3.2, H3.1, H3.2. For any $K_{\mathsf{max}} \in \mathbb{N}$, let K be an independent discrete r.v. drawn uniformly from $\{0, ..., K_{\mathsf{max}}-1\}$ and define the following quantity:*

$$\Delta_{(K_{\mathsf{max}})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})] + \sum_{k=0}^{K_{\mathsf{max}}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}}, \tag{3.3.5}$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}[\|\nabla\widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}, \quad \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}. \tag{3.3.6}$$

Proof The proof is postponed to Appendix 3.6

Next, we show that under an additional assumption on the sequence of batch size $M_{(k)}$, the MISSO method converges almost surely to a stationary point:

Theorem 2 *Under S3.1, S3.2, H3.1, H3.2. In addition, assume that $\{M_{(k)}\}_{k\geq 0}$ is a non-decreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$. Then:*

1. *the negative part of the stationarity measure converges almost surely to zero,* i.e., $\lim_{k\to\infty} g_-(\boldsymbol{\theta}^{(k)}) = 0$ *a.s.*.
2. *the objective value $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges almost surely to a finite number $\underline{\mathcal{L}}$,* i.e., $\lim_{k\to\infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$ *a.s.*.

Proof The proof is postponed to Appendix 3.7

In particular, the first result above shows that the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ produced by the MISSO method satisfies an *asymptotic stationary point condition*. Note that $\Delta_{(K_{\mathsf{max}})}$ is

finite for any $K_{\mathsf{max}} \in \mathbb{N}$. As expected, the MISSO method converges to a stationary point of (3.1.1) asymptotically and at a sublinear rate $\mathbb{E}[g_-^{(K)}] \leq -\mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$. Furthermore, we remark that the MISO method can be analyzed in 1 as a special case of the MISSO method satisfying $C_{\mathsf{r}} = C_{\mathsf{gr}} = 0$. In this case, while the asymptotic convergence is well known from [Mairal, 2015b] [cf. H3.2], Eq. (3.3.6) gives a non-asymptotic rate of $\mathbb{E}[g_-^{(K)}] \leq -\mathcal{O}(\sqrt{nL/K_{\mathsf{max}}})$ which is new to our best knowledge.

3.4 Application to Logistic Regression and Bayesian Deep Learning

3.4.1 Binary logistic regression with missing values

This application follows **Example 1** described in Section 3.2. We consider a binary regression setup, $((y_i, z_i), i \in [\![n]\!])$ where $y_i \in \{0, 1\}$ is a binary response and $z_i = (z_{i,j} \in \mathbb{R}, j \in [\![p]\!])$ is a covariate vector. The vector of covariates $z_i = [z_{i,\text{mis}}, z_{i,\text{obs}}]$ is not fully observed where we denote by $z_{i,\text{mis}}$ the missing values and $z_{i,\text{obs}}$ the observed covariate. It is assumed that $(z_i, i \in [\![n]\!])$ are i.i.d. and marginally distributed according to $\mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Omega})$ where $\beta \in \mathbb{R}^p$ and Ω is a positive definite $p \times p$ matrix.

We define the conditional distribution of the observations y_i given $z_i = (z_{i,\text{mis}}, z_{i,\text{obs}})$ as:

$$p_i(y_i|z_i) = S(\boldsymbol{\delta}^\top \bar{z}_i)^{y_i} \left(1 - S(\boldsymbol{\delta}^\top \bar{z}_i)\right)^{1-y_i} \tag{3.4.1}$$

where for $u \in \mathbb{R}$, $S(u) = 1/(1 + \mathrm{e}^{-u})$, $\boldsymbol{\delta} = (\delta_0, \cdots, \delta_p)$ are the logistic parameters and $\bar{z}_i = (1, z_i)$. We are interested in estimating $\boldsymbol{\delta}$ and finding the latent structure of the covariates z_i. Here, $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Omega})$ is the parameter to estimate. For $i \in [\![n]\!]$, the complete data log-likelihood is expressed as:

$$\log f_i(z_{i,\text{mis}}, \boldsymbol{\theta}) \propto y_i \boldsymbol{\delta}^\top \bar{z}_i - \log\left(1 + \exp(\boldsymbol{\delta}^\top \bar{z}_i)\right) - \frac{1}{2}\log(|\boldsymbol{\Omega}|) + \frac{1}{2}\mathrm{Tr}\left(\boldsymbol{\Omega}^{-1}(z_i - \boldsymbol{\beta})(z_i - \boldsymbol{\beta})^\top\right).$$

Choice of surrogate function for MISO: We recall the MISO deterministic surrogate defined in (3.2.8):

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) = \int_{\mathsf{Z}} \log\left(p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) / f_i(z_{i,\text{mis}}, \boldsymbol{\theta})\right) p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i) \,. \tag{3.4.2}$$

where $\boldsymbol{\theta} = (\delta, \beta, \Omega)$ and $\overline{\boldsymbol{\theta}} = (\bar{\delta}, \bar{\beta}, \bar{\Omega})$. We adapt it to our missing covariates problem and decompose the term depending on θ, while $\bar{\theta}$ is fixed, in two following parts:

$$\begin{aligned}
\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) &\approx -\int_{\mathsf{Z}} \log f_i(z_{i,\text{mis}}, z_{i,\text{obs}}, \boldsymbol{\theta})\, p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}}) \\
&= -\int_{\mathsf{Z}} \log \left[p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) p_i(z_{i,\text{mis}}, \beta, \Omega) \right] p_i(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}}) \\
&= \underbrace{-\int_{\mathsf{Z}} \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta)\, p_i(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}})}_{=\hat{\mathcal{L}}_i^{(1)}(\delta, \overline{\boldsymbol{\theta}})} \underbrace{-\int_{\mathsf{Z}} \log p_i(z_{i,\text{mis}}, \beta, \Omega)\, p_i(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}})}_{=\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \overline{\boldsymbol{\theta}})}
\end{aligned} \tag{3.4.3}$$

The mean β and the covariance Ω of the latent structure can be estimated minimizing the sum of MISSO surrogates $\tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M)$, defined as MC approximation of $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \overline{\boldsymbol{\theta}})$, for all $i \in [\![n]\!]$, in closed-form expression.

We thus keep the surrogate $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \overline{\boldsymbol{\theta}})$ and consider the following quadratic approximation of $\hat{\mathcal{L}}_i^{(1)}(\delta, \overline{\boldsymbol{\theta}})$ to estimate the vector of logistic parameters δ:

$$\begin{aligned}
&\hat{\mathcal{L}}_i^{(1)}(\bar{\delta}, \overline{\boldsymbol{\theta}}) - \int_{\mathsf{Z}} \nabla \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta)\big|_{\delta=\bar{\delta}}\, p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}})(\delta - \bar{\delta}) \\
&\quad - (\delta - \bar{\delta})/2 \int_{\mathsf{Z}} \nabla^2 \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta)\, p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}})\, p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}})(\delta - \bar{\delta})^\top
\end{aligned} \tag{3.4.4}$$

Recall that:

$$\begin{aligned}
\nabla \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) &= z_i \left(y_i - S(\delta^\top z_i) \right) \\
\nabla^2 \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) &= -z_i z_i^\top \dot{S}(\delta^\top z_i)
\end{aligned} \tag{3.4.5}$$

where $\dot{S}(u)$ is the derivative of $S(u)$. Note that $\dot{S}(u) \leq 1/4$ and since, for all $i \in [\![n]\!]$, the $p \times p$ matrix $z_i z_i^\top$ is semi-definite positive we can assume:

L1 *For all $i \in [\![n]\!]$ and $\epsilon > 0$, there exist, for all $z_i \in \mathsf{Z}$, a positive definite matrix $H_i(z_i) := \frac{1}{4}(z_i z_i^\top + \epsilon I_d)$ such that for all $\delta \in \mathbb{R}^p$, $-z_i z_i^\top \dot{S}(\delta^\top z_i) \leq H_i(z_i)$.*

We thus use, for all $i \in [\![n]\!]$, the following surrogate function to estimate δ:

$$\bar{\mathcal{L}}_i^{(1)}(\delta, \overline{\boldsymbol{\theta}}) = \hat{\mathcal{L}}_i^{(1)}(\bar{\delta}, \overline{\boldsymbol{\theta}}) - D_i^\top (\delta - \bar{\delta}) + \frac{1}{2}(\delta - \bar{\delta}) H_i (\delta - \bar{\delta})^\top \tag{3.4.6}$$

where:

$$\begin{aligned}
D_i &= \int_{\mathsf{Z}} \nabla \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta)\big|_{\delta=\bar{\delta}}\, p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}}) \\
H_i &= \int_{\mathsf{Z}} H_i(z_{i,\text{mis}})\, p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}})
\end{aligned} \tag{3.4.7}$$

Finally, at iteration k, the total surrogate is:

$$\begin{aligned}\tilde{\mathcal{L}}^{(k)}(\theta) &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathcal{L}}_i(\theta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) \\ &= \frac{1}{n}\sum_{i=1}^{n}\tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) - \frac{1}{n}\sum_{i=1}^{n}\tilde{D}_i^{(\tau_i^k)}(\delta - \delta^{(\tau_i^k)}) \\ &+ \frac{1}{2n}\sum_{i=1}^{n}(\delta - \delta^{(\tau_i^k)})\left\{\tilde{H}_i^{(\tau_i^k)}\right\}(\delta - \delta^{(\tau_i^k)})^\top\end{aligned} \tag{3.4.8}$$

where for all $i \in [\![n]\!]$:

$$\begin{aligned}\tilde{D}_i^{(\tau_i^k)} &= \frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(\tau_i^k)}\left(y_i - S(\left(\delta^{(\tau_i^k)}\right)^\top z_{i,m}(\tau_i^k))\right) \\ \tilde{H}_i^{(\tau_i^k)} &= \frac{1}{4M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(\tau_i^k)}(z_{i,m}^{(\tau_i^k)})^\top\end{aligned} \tag{3.4.9}$$

Minimizing the total surrogate (3.4.8) boils down to performing a quasi-Newton step. It is perhaps sensible to apply some diagonal loading which is perfectly compatible with the surrogate interpretation we just gave.

MISSO update: At the k-th iteration, and after the initialization, for all $i \in [\![n]\!]$, of the latent variables $(z_i^{(0)})$, the MISSO algorithm consists in picking an index i_k uniformly on $[\![n]\!]$, completing the observations by sampling a Monte Carlo batch $\{z_{i_k,\text{mis},m}^{(k)}\}_{m=1}^{M_{(k)}}$ of missing values from the conditional distribution $p(z_{i_k,\text{mis}}|z_{i_k,y_{1:N}}, y_{i_k}; \boldsymbol{\theta}^{(k-1)})$ using an MCMC sampler and computing the estimated parameters as follows:

$$\begin{aligned}\boldsymbol{\beta}^{(k)} &= \arg\min_{\beta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega^{(k)}, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(k)} \\ \boldsymbol{\Omega}^{(k)} &= \arg\min_{\Omega\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathcal{L}}_i^{(2)}(\beta^{(k)}, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(k)}(z_{i,m}^{(k)})^\top - \boldsymbol{\beta}^{(k)}(\boldsymbol{\beta}^{(k)})^\top .\end{aligned} \tag{3.4.10}$$

where $z_{i,m}^{(k)} = (z_{i,\text{mis},m}^{(k)}, z_{i,y_{1:N}})$ is composed of a simulated and an observed part. The logistic parameters are estimated as follows:

$$\boldsymbol{\delta}^{(k)} = \arg\min_{\delta\in\Theta}\frac{1}{n}\sum_{i=1}^{n}\tilde{\mathcal{L}}_i^{(1)}(\delta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) \tag{3.4.11}$$

where $\tilde{\mathcal{L}}_i^{(1)}(\delta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}})$ is the MC approximation of the MISO surrogate defined in (3.4.6)and which leads to the following quasi-Newton step:

$$\boldsymbol{\delta}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \boldsymbol{\delta}^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1}\tilde{D}^{(k)} \tag{3.4.12}$$

with $\tilde{D}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{D}_i^{(\tau_i^k)}$ and $\tilde{H}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \tilde{H}_i^{(\tau_i^k)}$.

Fitting a logistic regression model on the TraumaBase dataset We apply the MISSO method to fit a logistic regression model on the TraumaBase (`http://traumabase.eu`) dataset, which consists of data collected from 15 trauma centers in France, covering measurements on patients from the initial to last stage of trauma.

Similar to [Jiang et al., 2018], we select $p = 16$ influential quantitative measurements, described in Appendix 3.9.1, on $n = 6384$ patients, and we adopt the logistic regression model with missing covariates in (3.4.1) to predict the risk of a severe hemorrhage which is one of the main cause of death after a major trauma. Note as the dataset considered is heterogeneous – coming from multiple sources with frequently missed entries – we apply the latent data model described in the above. For the Monte-Carlo sampling of $z_{i,mis}$, we run a Metropolis Hastings algorithm with the target distribution $p(\cdot|z_{i,obs}, y_i; \boldsymbol{\theta}^{(k)})$ whose procedure is detailed in Appendix 3.9.1.

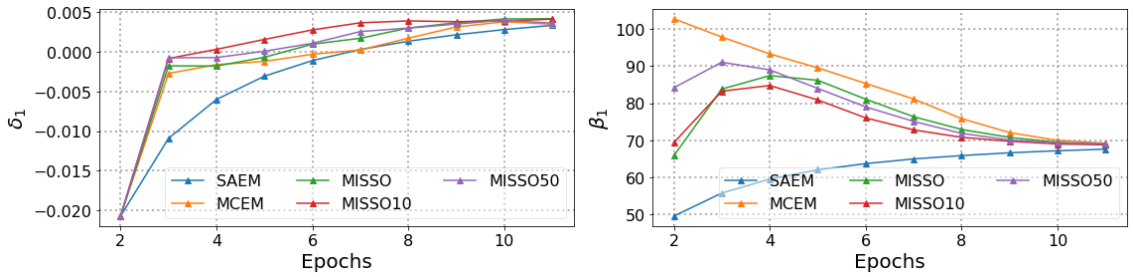

Figure 3.1 – Convergence of first component of the vector of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against the number of passes over the data.

We compare in Figure 3.1 the convergence behavior of the estimated parameters $\boldsymbol{\beta}$ using SAEM [Delyon et al., 1999a] (with stepsize $\gamma_k = 1/k$), MCEM [Wei and Tanner, 1990a] and the proposed MISSO method. For the MISSO method, we set the batch size to $M_{(k)} = 10 + k^2$ and we examine with selecting different number of functions in Line 5 in the method – the default settings with 1 function (MISSO), 10% (MISSO10) and 50% (MISSO50) of the functions per iteration. From Figure 3.1, the MISSO method converges to a static value with less number of epochs than the MCEM, SAEM methods. It is worth noting that the difference among the MISSO runs for different number of selected functions demonstrates a variance-cost tradeoff.

3.4.2 Fitting Bayesian LeNet-5 on MNIST

This application follows **Example 2** described in Section 3.2. We apply the MISSO method to fit a Bayesian variant of LeNet-5 [LeCun et al., 1998] (see Appendix 3.9.2). We train this network on the MNIST dataset [LeCun, 1998]. The training set is composed of $N = 55\,000$ handwritten digits, 28×28 images. Each image is labelled with its corresponding number (from zero to nine). Under the prior distribution π, see (3.2.9), the weights are assumed independent and identically distributed according to $\mathcal{N}(0, 1)$. We also assume that $q(\cdot; \boldsymbol{\theta}) \equiv \mathcal{N}(\mu, \sigma^2 \mathbf{I})$. The variational posterior parameters are thus $\boldsymbol{\theta} = (\mu, \sigma)$ where $\mu = (\mu_\ell, \ell \in [\![d]\!])$ where d is the number of weights in the neural network. We use the re-parametrization as $w = t(\boldsymbol{\theta}, z) = \mu + \sigma z$ with $z \sim \mathcal{N}(0, \mathbf{I})$.

At iteration k, minimizing the sum of stochastic surrogates defined as in (3.2.4) and (3.2.14) yields the following MISSO update — step (i) pick a function index i_k uniformly on $[\![n]\!]$; step (ii) sample a Monte Carlo batch $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ from $\mathcal{N}(0, \mathbf{I})$; and step (iii) update the parameters as

$$\mu_\ell^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\mu_\ell, i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n} \sum_{i=1}^{n} \sigma^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\sigma, i}^{(k)} \,, \tag{3.4.13}$$

where $\hat{\boldsymbol{\delta}}_{\mu_\ell, i}^{(k)} = \hat{\boldsymbol{\delta}}_{\mu_\ell, i}^{(k-1)}$ and $\hat{\boldsymbol{\delta}}_{\sigma, i}^{(k)} = \hat{\boldsymbol{\delta}}_{\sigma, i}^{(k-1)}$ for $i \neq i_k$ and:

$$\hat{\boldsymbol{\delta}}_{\mu_\ell, i_k}^{(k)} = -\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \nabla_w \log p(y_{i_k} | x_{i_k}, w) \Big|_{w = t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} d(\boldsymbol{\theta}^{(k-1)}) \,,$$

$$\hat{\boldsymbol{\delta}}_{\sigma, i_k}^{(k)} = -\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} z_m^{(k)} \nabla_w \log p(y_{i_k} | x_{i_k}, w) \Big|_{w = t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\sigma} d(\boldsymbol{\theta}^{(k-1)})$$

with $d(\boldsymbol{\theta}) = n^{-1} \sum_{\ell=1}^{d} \left(-\log(\sigma) + (\sigma^2 + \mu_\ell^2)/2 - 1/2 \right)$.

We compare the convergence of the *Monte Carlo variants* of the following state of the art optimization algorithms — the ADAM [Kingma and Ba, 2015], the Momentum [Sutskever et al., 2013] and the SAG [Schmidt et al., 2017] methods versus the *Bayes by Backprop* (BBB) [Blundell et al., 2015] and our proposed MISSO method. For all these methods, the loss function (3.2.11) and its gradients were computed by Monte Carlo integration using Tensorflow Probability library [Dillon et al., 2017], based on the re-parametrization described above. Update rules for each algorithm are performed using their vanilla implementations on TensorFlow [Abadi et al., 2015] as detailed in Appendix 3.9.2. We use the following hyperparameters for all runs — the learning rate is 10^{-3}, we run 100 epochs with a mini-batch size of 128 and use the batchsize of $M_{(k)} = k$.

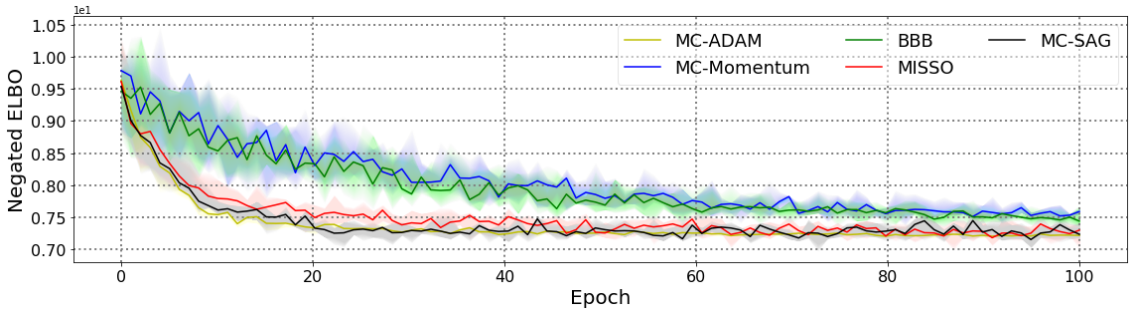

Figure 3.2 – (Incremental Variational Inference) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

Figure 3.2 shows the convergence of the negated evidence lower bound against the number of passes over data (one pass represents an epoch). As observed, the proposed MISSO method outperforms *Bayes by Backprop* and Momentum, while similar convergence rates are observed with the MISSO, ADAM and SAG methods.

3.5 Conclusions

We presented a unifying framework for minimizing a non-convex finite-sum objective function using incremental surrogates when the latter functions are expressed as an expectation and are intractable. Our approach covers a large class of non-convex applications in machine learning such as logistic regression with missing values and variational inference. We provide both finite-time and asymptotic guarantees of our incremental stochastic surrogate optimization technique and illustrate our findings training a binary logistic regression with missing covariates to predict hemorrhagic shock and a Bayesian variant of LeNet-5 on MNIST.

Appendices to Incr. Method for Non-smooth Non-convex Optimization

3.6 Proof of Theorem 1

Theorem *Under S3.1, S3.2, H3.1, H3.2. For any $K_{\mathsf{max}} \in \mathbb{N}$, let K be an independent discrete r.v. drawn uniformly from $\{0, ..., K_{\mathsf{max}} - 1\}$ and define the following quantity:*

$$\Delta_{(K_{\mathsf{max}})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})] + \sum_{k=0}^{K_{\mathsf{max}}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}},$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}, \quad \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}.$$

Proof We begin by recalling the definition

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta}). \tag{3.6.1}$$

Notice that

$$\begin{aligned}\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^{k+1})}, \{z_{i,m}^{(\tau_i^{k+1})}\}_{m=1}^{M_{(\tau_i^{k+1})}}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n}(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})).\end{aligned} \tag{3.6.2}$$

Furthermore, we recall that

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n}\textstyle\sum_{i=1}^{n} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \qquad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{3.6.3}$$

Due to S3.2, we have

$$\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{3.6.4}$$

To prove the first bound in (3.3.6), using the optimality of $\boldsymbol{\theta}^{(k+1)}$, one has

$$\begin{aligned}
&\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\
&= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big)
\end{aligned} \tag{3.6.5}$$

Let $\mathcal{F}_k$ be the filtration of random variables $\{i_{k-1},\{z_{i_{k-1},m}^{(k-1)}\}_{m=1}^{M_{(k-1)}},\boldsymbol{\theta}^{(k)}\}$ up to iteration k. We observe that the conditional expectation evaluates to

$$\begin{aligned}
&\mathbb{E}_{i_k}\big[\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}})|\mathcal{F}_k,i_k\big]|\mathcal{F}_k\big] \\
&= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathbb{E}_{i_k}\big[\mathbb{E}\big[\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} r_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},z_{i_k,m}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})|\mathcal{F}_k,i_k\big]|\mathcal{F}_k\big] \\
&\leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}},
\end{aligned} \tag{3.6.6}$$

where the last inequality is due to H3.2. Moreover,

$$\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})|\mathcal{F}_k\big] = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)},\{z_{i,m}^{(\tau_i^k)}\}_{m=1}^{M_{(\tau_i^k)}}) = \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{3.6.7}$$

Taking the conditional expectations on both sides of (3.6.5) and re-arranging terms give:

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\,\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} \tag{3.6.8}$$

Proceeding from (3.6.8), we observe the following lower bound for the left hand side

$$\begin{aligned}
&\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \overset{(a)}{=} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \\
&\overset{(b)}{\geq} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2L}\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} r_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)},z_{i,m}^{(\tau_i^k)}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)})\Big\}}_{:=-\delta^{(k)}(\boldsymbol{\theta}^{(k)})} + \frac{1}{2L}\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2
\end{aligned} \tag{3.6.9}$$

where (a) is due to $\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$ [cf. S3.1], (b) is due to (3.6.4) and we have defined the

summation in the last equality as $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$. Substituting the above into (3.6.8) yields

$$\frac{\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \leq n\,\mathbb{E}[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)}) \tag{3.6.10}$$

Observe the following upper bound on the total expectations:

$$\mathbb{E}[\delta^{(k)}(\boldsymbol{\theta}^{(k)})] \leq \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{C_{\mathsf{r}}}{\sqrt{M_{(\tau_i^k)}}}\Big], \tag{3.6.11}$$

which is due to H3.2. It yields

$$\mathbb{E}[\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \leq 2nL\,\mathbb{E}[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})] + \frac{2LC_{\mathsf{r}}}{\sqrt{M_{(k)}}} + \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}\Big[\frac{2LC_{\mathsf{r}}}{\sqrt{M_{(\tau_i^k)}}}\Big]$$

Finally, for any $K_{\mathsf{max}} \in \mathbb{N}$, we let K be a discrete r.v. that is uniformly drawn from $\{0, 1, ..., K_{\mathsf{max}} - 1\}$. Using H3.2 and taking total expectations lead to

$$\begin{aligned}
&\mathbb{E}[\|\nabla\widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] = \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{2LC_{\mathsf{r}}}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}\Big[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{M_{(\tau_i^k)}}}\Big]
\end{aligned} \tag{3.6.12}$$

For all $i \in [\![1, n]\!]$, the index i is selected with a probability equal to $\frac{1}{n}$ when conditioned independently on the past. We observe:

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^{k}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} \tag{3.6.13}$$

Taking the sum yields:

$$\begin{aligned}
&\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{j=1}^{k}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{l=0}^{k-1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{k-(l+1)} M_{(l)}^{-1/2} \\
&= \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}\sum_{k=l+1}^{K_{\mathsf{max}}-1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{k-(l+1)} \leq \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}
\end{aligned} \tag{3.6.14}$$

where the last inequality is due to upper bounding the geometric series. Plugging this

back into (3.6.12) yields

$$\begin{aligned}\mathbb{E}[\|\nabla\widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] &= \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\ &\leq \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}} = \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}.\end{aligned} \tag{3.6.15}$$

This concludes our proof for the first inequality in (3.3.6). To prove the second inequality of (3.3.6), we define the shorthand notations $g^{(k)} := g(\boldsymbol{\theta}^{(k)})$, $g_-^{(k)} := -\min\{0, g^{(k)}\}$, $g_+^{(k)} := \max\{0, g^{(k)}\}$. We observe that

$$\begin{aligned}g^{(k)} &= \inf_{\boldsymbol{\theta}\in\Theta}\frac{\mathcal{L}'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)})}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|} \\ &= \inf_{\boldsymbol{\theta}\in\Theta}\Big\{\frac{\frac{1}{n}\sum_{i=1}^n\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|} - \frac{\left\langle\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \,|\, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}\right\rangle}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|}\Big\} \\ &\geq -\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta}\frac{\frac{1}{n}\sum_{i=1}^n\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|}\end{aligned} \tag{3.6.16}$$

where the last inequality is due to the Cauchy-Schwarz inequality and we have defined $\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}, \boldsymbol{d}; \boldsymbol{\theta}^{(\tau_i^k)})$ as the directional derivative of $\widehat{\mathcal{L}}_i(\cdot; \boldsymbol{\theta}^{(\tau_i^k)})$ at $\boldsymbol{\theta}$ along the direction $\boldsymbol{d}$. Moreover, for any $\boldsymbol{\theta} \in \Theta$,

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^n\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\ &= \underbrace{\widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)})}_{\geq 0} - \widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}) + \frac{1}{n}\sum_{i=1}^n\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\ &\geq \frac{1}{n}\sum_{i=1}^n\Big\{\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) - \frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} r_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)})\Big\}\end{aligned} \tag{3.6.17}$$

where the inequality is due to the optimality of $\boldsymbol{\theta}^{(k)}$ and the convexity of $\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta})$ [cf. H3.1]. Denoting a scaled version of the above term as:

$$\epsilon^{(k)}(\boldsymbol{\theta}) := \frac{\frac{1}{n}\sum_{i=1}^n\Big\{\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} r_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) - \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})\Big\}}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|}.$$

We have

$$g^{(k)} \geq -\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta}(-\epsilon^{(k)}(\boldsymbol{\theta})) \geq -\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| - \sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{3.6.18}$$

Since $g^{(k)} = g_+^{(k)} - g_-^{(k)}$ and $g_+^{(k)} g_-^{(k)} = 0$, this implies

$$g_-^{(k)} \leq \|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \sup_{\boldsymbol{\theta}\in\Theta} |\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{3.6.19}$$

Consider the above inequality when $k = K$, *i.e.,* the random index, and taking total expectations on both sides gives

$$\mathbb{E}[g_-^{(K)}] \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|] + \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(K)}(\boldsymbol{\theta})] \tag{3.6.20}$$

We note that

$$\Big(\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|]\Big)^2 \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta(K_{\mathsf{max}})}{K_{\mathsf{max}}}, \tag{3.6.21}$$

where the first inequality is due to the convexity of $(\cdot)^2$ and the Jensen's inequality, and

$$\begin{aligned}\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(K)}(\boldsymbol{\theta})] = \frac{1}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}} \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(k)}(\boldsymbol{\theta})] &\overset{(a)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} M_{(\tau_i^k)}^{-1/2}\Big] \\ &\overset{(b)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}\end{aligned} \tag{3.6.22}$$

where (a) is due to H3.2 and (b) is due to (3.6.14). This implies

$$\mathbb{E}[g_-^{(K)}] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}, \tag{3.6.23}$$

and concludes the proof of the theorem. ■

3.7 Proof of Theorem 2

Theorem *Under S3.1, S3.2, H3.1, H3.2. In addition, assume that* $\{M_{(k)}\}_{k\geq 0}$ *is a non-decreasing sequence of integers which satisfies* $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$. *Then:*

1. *the negative part of the stationarity measure converges almost surely to zero,* i.e., $\lim_{k\to\infty} g_-(\boldsymbol{\theta}^{(k)}) = 0$ *a.s..*
2. *the objective value* $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ *converges almost surely to a finite number* $\underline{\mathcal{L}}$, i.e., $\lim_{k\to\infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$ *a.s..*

Proof We apply the following auxiliary lemma which proof can be found in Appendix 3.8 for the readability of the current proof:

Lemma 1 *Let $(V_k)_{k\geq 0}$ be a non negative sequence of random variables such that $\mathbb{E}[V_0] < \infty$. Let $(X_k)_{k\geq 0}$ a non negative sequence of random variables and $(E_k)_{k\geq 0}$ be a sequence of random variables such that $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|] < \infty$. If for any $k \geq 1$:*

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1} \tag{3.7.1}$$

then:

1. *for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.*
2. *the sequence $(\mathbb{E}[V_k])_{k\geq 0}$ converges and $\lim_{k\to\infty} \mathbb{E}[V_k] = \mathbb{E}[V_\infty]$.*
3. *the series $\sum_{k=0}^{\infty} X_k$ converges almost surely and $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.*

We proceed from (3.6.5) by re-arranging terms and observing that

$$\begin{aligned}
\widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq\ & \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tfrac{1}{n}(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\
& - (\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\
& + \tfrac{1}{n}(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\
& + \tfrac{1}{n}(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}}))
\end{aligned} \tag{3.7.2}$$

Our idea is to apply Lemma 1. Under S3.1, the finite sum of surrogate functions $\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta})$, defined in (3.3.4), is lower bounded by a constant $c_k > -\infty$ for any $\boldsymbol{\theta}$. To this end, we observe that

$$V_k := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \inf_{k\geq 0} c_k \geq 0 \tag{3.7.3}$$

is a non-negative random variable.

Secondly, under H3.1, the following random variable is non-negative

$$X_k := \tfrac{1}{n}(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(\tau_{i_k}^k)}; \boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \geq 0. \tag{3.7.4}$$

Thirdly, we define

$$\begin{aligned}
E_k =\ & -(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\
& + \tfrac{1}{n}(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\
& + \tfrac{1}{n}(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})).
\end{aligned} \tag{3.7.5}$$

Note that from the definitions (3.7.3), (3.7.4), (3.7.5), we have $V_{k+1} \leq V_k - X_k + E_k$ for any $k \geq 1$.

Under H3.2, we observe that

$$\mathbb{E}\big[|\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}})-\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})|\big]\leq C_{\mathsf{r}}M_{(k)}^{-1/2} \tag{3.7.6}$$

$$\mathbb{E}\Big[\Big|\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)})-\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\Big|\Big]\leq C_{\mathsf{r}}\mathbb{E}\Big[M_{(\tau_{i_k}^k)}^{-1/2}\Big] \tag{3.7.7}$$

$$\mathbb{E}\big[|\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})-\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})|\big]\leq \tfrac{1}{n}\textstyle\sum_{i=1}^n C_{\mathsf{r}}\mathbb{E}\Big[M_{(\tau_i^k)}^{-1/2}\Big] \tag{3.7.8}$$

Therefore,

$$\mathbb{E}\big[|E_k|\big]\leq \tfrac{C_{\mathsf{r}}}{n}\Big(M_{(k)}^{-1/2}+\mathbb{E}\Big[M_{(\tau_{i_k}^k)}^{-1/2}+\textstyle\sum_{i=1}^n\big\{M_{(\tau_i^k)}^{-1/2}+M_{(\tau_i^{k+1})}^{-1/2}\big\}\Big]\Big) \tag{3.7.9}$$

Using (3.6.14) and the assumption on the sequence $\{M_{(k)}\}_{k\geq 0}$, we obtain that

$$\sum_{k=0}^{\infty}\mathbb{E}\big[|E_k|\big]<\frac{C_{\mathsf{r}}}{n}(2+2n)\sum_{k=0}^{\infty}M_{(k)}^{-1/2}<\infty. \tag{3.7.10}$$

Therefore, the conclusions in Lemma 1 hold. Precisely, we have $\sum_{k=0}^{\infty}X_k<\infty$ and $\sum_{k=0}^{\infty}\mathbb{E}[X_k]<\infty$ almost surely. Note that this implies

$$\begin{aligned}\infty>\sum_{k=0}^{\infty}\mathbb{E}[X_k]&=\frac{1}{n}\sum_{k=0}^{\infty}\mathbb{E}\big[\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)})-\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\big]\\&=\frac{1}{n}\sum_{k=0}^{\infty}\mathbb{E}\big[\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})-\mathcal{L}(\boldsymbol{\theta}^{(k)})\big]=\frac{1}{n}\sum_{k=0}^{\infty}\mathbb{E}\big[\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\big]\end{aligned} \tag{3.7.11}$$

Since $\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\geq 0$, the above implies

$$\lim_{k\to\infty}\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})=0\quad\text{a.s.} \tag{3.7.12}$$

and subsequently applying (3.6.4), we have $\lim_{k\to\infty}\|\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|=0$ almost surely. Finally, it follows from (3.6.4) and (3.6.19) that

$$\lim_{k\to\infty}g_-^{(k)}\leq\lim_{k\to\infty}\sqrt{2L}\sqrt{\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})}+\lim_{k\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|=0, \tag{3.7.13}$$

where the last equality holds almost surely due to the fact that $\sum_{k=0}^{\infty}\mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|]<\infty$. This concludes the asymptotic convergence of the MISSO method.

Finally, we prove that $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges almost surely. As a consequence of Lemma 1, it is clear that $\{V_k\}_{k\geq 0}$ converges almost surely and so is $\{\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\}_{k\geq 0}$, *i.e.,* we have $\lim_{k\to\infty}\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})=\underline{\mathcal{L}}$. Applying (3.7.12) implies that

$$\underline{\mathcal{L}}=\lim_{k\to\infty}\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})=\lim_{k\to\infty}\mathcal{L}(\boldsymbol{\theta}^{(k)})\quad\text{a.s.} \tag{3.7.14}$$

This shows that $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges almost surely to $\underline{\mathcal{L}}$. ■

3.8 Proof of Lemma 1

Lemma *Let $(V_k)_{k\geq 0}$ be a non negative sequence of random variables such that $\mathbb{E}[V_0] < \infty$. Let $(X_k)_{k\geq 0}$ a non negative sequence of random variables and $(E_k)_{k\geq 0}$ be a sequence of random variables such that $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|] < \infty$. If for any $k \geq 1$:*

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1}$$

then:

1. *for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.*
2. *the sequence $(\mathbb{E}[V_k])_{k\geq 0}$ converges and $\lim_{k\to\infty} \mathbb{E}[V_k] = \mathbb{E}[V_\infty]$.*
3. *the series $\sum_{k=0}^{\infty} X_k$ converges almost surely and $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.*

Proof We first show that for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$. Note indeed that:

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \tag{3.8.1}$$

showing that $\mathbb{E}[V_k] \leq \mathbb{E}[V_0] + \mathbb{E}\left[\sum_{j=1}^{k} E_j\right] < \infty$.

Since $0 \leq X_k \leq V_{k-1} - V_k + E_k$ we also obtain for all $k \geq 0$, $\mathbb{E}[X_k] < \infty$. Moreover, since $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$, the series $\sum_{j=1}^{\infty} E_j$ converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \tag{3.8.2}$$

Note that $\mathbb{E}[|W_k|] \leq \mathbb{E}[V_k] + \mathbb{E}\left[\sum_{j=k+1}^{\infty} |E_j|\right] < \infty$. For all $k \geq 1$, we get:

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \end{aligned} \tag{3.8.3}$$

Hence the sequences $(W_k)_{k\geq 0}$ and $(\mathbb{E}[W_k])_{k\geq 0}$ are non increasing. Since for all $k \geq 0$, $W_k \geq -\sum_{j=1}^{\infty} |E_j| > -\infty$ and $\mathbb{E}[W_k] \geq -\sum_{j=1}^{\infty} \mathbb{E}[|E_j|] > -\infty$, the (random) sequence $(W_k)_{k\geq 0}$ converges a.s. to a limit W_∞ and the (deterministic) sequence $(\mathbb{E}[W_k])_{k\geq 0}$ converges to a

limit w_∞. Since $|W_k| \leq V_0 + \sum_{j=1}^{\infty} |E_j|$, the Fatou lemma implies that:

$$\mathbb{E}[\liminf_{k\to\infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k\to\infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \tag{3.8.4}$$

showing that the random variable W_∞ is integrable.

In the sequel, set $U_k \triangleq W_0 - W_k$. By construction we have for all $k \geq 0$, $U_k \geq 0$, $U_k \leq U_{k+1}$ and $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$ and by the monotone convergence theorem, we get:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k\to\infty} U_k] \tag{3.8.5}$$

Finally, we have:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k\to\infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \tag{3.8.6}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (3.8.3) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\begin{aligned} &\sum_{j=1}^{\infty} X_j \leq W_0 - W_\infty < \infty \\ &\sum_{j=1}^{\infty} \mathbb{E}[X_j] \leq \mathbb{E}[W_0] - w_\infty < \infty \end{aligned} \tag{3.8.7}$$

which concludes the proof of the lemma. ■

3.9 Details about the Numerical Experiments

3.9.1 Binary Logistic Regression on the Traumabase

Traumabase quantitative variables The list of the 16 quantitative variables we use in our experiments are as follows — *age, weight, height, BMI (Body Mass Index), the Glasgow Coma Scale, the Glasgow Coma Scale motor component, the minimum systolic blood pressure, the minimum diastolic blood pressure, the maximum number of heart rate (or pulse) per unit time (usually a minute), the systolic blood pressure at arrival of ambulance, the diastolic blood pressure at arrival of ambulance, the heart rate at arrival of ambulance, the capillary Hemoglobin concentration, the oxygen saturation, the fluid expansion colloids, the fluid expansion cristalloids, the pulse pressure for the minimum value of diastolic and systolic blood pressure, the pulse pressure at arrival of ambulance.*

Algorithm 3.3 MH aglorithm

Input: initialization $z_{i,mis,0} \sim q(z_{i,mis}; \boldsymbol{\delta})$
for $m = 1, \cdots, M$ **do**
  Sample $z_{i,mis,m} \sim q(z_{i,mis}; \boldsymbol{\delta})$
  Sample $u \sim \mathcal{U}([\![0, 1]\!])$
  Calculate the ratio $r = \frac{\pi(z_{i,mis,m};\boldsymbol{\theta})/q(z_{i,mis,m});\boldsymbol{\delta})}{\pi(z_{i,mis,m-1};\boldsymbol{\theta})/q(z_{i,mis,m-1});\boldsymbol{\delta})}$
  if $u < r$ **then**
    Accept $z_{i,mis,m}$
  else
    $z_{i,mis,m} \leftarrow z_{i,mis,m-1}$
  end if
end for
Output: $z_{i,mis,M}$

Metropolis Hastings algorithm During the simulation step of the MISSO method, the sampling from the target distribution $\pi(z_{i,mis}; \boldsymbol{\theta}) := p(z_{i,mis}|z_{i,obs}, y_i; \boldsymbol{\theta})$ is performed using a Metropolis Hastings (MH) algorithm [Meyn and Tweedie, 2012] with proposal distribution $q(z_{i,mis}; \boldsymbol{\delta}) := p(z_{i,mis}|z_{i,obs}; \boldsymbol{\delta})$ where $\boldsymbol{\theta} = (\beta, \Omega)$ and $\boldsymbol{\delta} = (\xi, \Sigma)$. The parameters of the Gaussian conditional distribution of $z_{i,mis}|z_{i,obs}$ read:

$$\begin{aligned} \xi &= \beta_{miss} + \Omega_{mis,obs}\Omega^{-1}_{obs,obs}(z_{i,obs} - \beta_{obs}) \ , \\ \Sigma &= \Omega_{mis,mis} + \Omega_{mis,obs}\Omega^{-1}_{obs,obs}\Omega_{obs,mis} \end{aligned} \tag{3.9.1}$$

where we have used the Schur Complement of $\Omega_{obs,obs}$ in Ω and noted β_{mis} (resp. β_{obs}) the missing (resp. observed) elements of β. The MH algorithm is summarized in 6.1.

3.9.2 Incremental Variational Inference for MNIST

Bayesian LeNet-5 Architecture We describe in Table 3.1 the architecture of the Convolutional Neural Network introduced in [LeCun et al., 1998] and trained on MNIST:

layer type	width	stride	padding	input shape	nonlinearity
convolution (5×5)	6	1	0	$1 \times 32 \times 32$	ReLU
max-pooling (2×2)		2	0	$6 \times 28 \times 28$	
convolution (5×5)	6	1	0	$1 \times 14 \times 14$	ReLU
max-pooling (2×2)		2	0	$16 \times 10 \times 10$	
fully-connected	120			400	ReLU
fully-connected	84			120	ReLU
fully-connected	10			84	

Table 3.1 – LeNet-5 architecture

Algorithms updates First, we initialize the means $\mu_\ell^{(0)}$ for $\ell \in [\![d]\!]$ and variance estimates $\sigma^{(0)}$. In the sequel, at iteration k and for all $i \in [\![n]\!]$ we define the following drift

terms:

$$\begin{aligned}\hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} &= -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} \nabla_w \log p(y_i|x_i,w)\Big|_{w=t(\boldsymbol{\theta}^{(k-1)},z_m^{(k)})} + \nabla_{\mu_\ell} d(\boldsymbol{\theta}^{(k-1)}) \;,\\ \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} &= -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} z_m^{(k)} \nabla_w \log p(y_i|x_i,w)\Big|_{w=t(\boldsymbol{\theta}^{(k-1)},z_m^{(k)})} + \nabla_{\sigma} d(\boldsymbol{\theta}^{(k-1)}) \;.\end{aligned} \tag{3.9.2}$$

For all benchmark algorithms, we pick, at iteration k, a function index i_k uniformly on $[\![n]\!]$ and sample a Monte Carlo batch $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ from the standard Gaussian distribution. The updates of the parameters μ_ℓ for all $\ell \in [\![d]\!]$ and σ break down as follows:

Monte Carlo SAG update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} \;, \tag{3.9.3}$$

where $\hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} = \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k-1)}$ and $\hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} = \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k-1)}$ for $i \neq i_k$ and are defined by (3.9.2) for $i = i_k$. The learning rate is set to $\gamma = 10^{-3}$.

Bayes By Backprop update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n}\hat{\boldsymbol{\delta}}_{\mu_\ell,i_k}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n}\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)} \;, \tag{3.9.4}$$

where the learning rate $\gamma = 10^{-3}$.

Monte Carlo Momentum update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} + \hat{\boldsymbol{v}}_{\mu_\ell}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} + \hat{\boldsymbol{v}}_{\sigma}^{(k)} \;, \tag{3.9.5}$$

where

$$\hat{\boldsymbol{v}}_{\mu_\ell,i}^{(k)} = \alpha\hat{\boldsymbol{v}}_{\mu_\ell}^{(k-1)} - \frac{\gamma}{n}\hat{\boldsymbol{\delta}}_{\mu_\ell,i_k}^{(k)} \quad \text{and} \quad \hat{\boldsymbol{v}}_{\sigma}^{(k)} = \alpha\hat{\boldsymbol{v}}_{\sigma}^{(k-1)} - \frac{\gamma}{n}\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)} \;, \tag{3.9.6}$$

where α and γ, respectively the momentum and the learning rates, are set to 10^{-3}.

Monte Carlo ADAM update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n}\hat{\boldsymbol{m}}_{\mu_\ell}^{(k)}/(\sqrt{\hat{\boldsymbol{m}}_{\mu_\ell}^{(k)}} + \epsilon) \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n}\hat{\boldsymbol{m}}_{\sigma}^{(k)}/(\sqrt{\hat{\boldsymbol{m}}_{\sigma}^{(k)}} + \epsilon) \;, \tag{3.9.7}$$

where

$$\begin{aligned}\hat{\boldsymbol{m}}_{\mu_\ell}^{(k)} &= \boldsymbol{m}_{\mu_\ell}^{(k-1)}/(1-\rho_1^k) \quad \text{with} \quad \boldsymbol{m}_{\mu_\ell}^{(k)} = \rho_1 \boldsymbol{m}_{\mu_\ell}^{(k-1)} + (1-\rho_1)\hat{\boldsymbol{\delta}}_{\mu_\ell,i_k}^{(k)} \;,\\ \hat{\boldsymbol{v}}_{\mu_\ell}^{(k)} &= \boldsymbol{v}_{\mu_\ell}^{(k-1)}/(1-\rho_2^k) \quad \text{with} \quad \boldsymbol{v}_{\mu_\ell}^{(k)} = \rho_2 \boldsymbol{v}_{\mu_\ell}^{(k-1)} + (1-\rho_1)(\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)})^2\end{aligned} \tag{3.9.8}$$

and

$$\begin{aligned}\hat{\boldsymbol{m}}_\sigma^{(k)} &= \boldsymbol{m}_\sigma^{(k-1)}/(1-\rho_1^k) \quad \text{with} \quad \boldsymbol{m}_\sigma^{(k)} = \rho_1 \boldsymbol{m}_\sigma^{(k-1)} + (1-\rho_1)\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)} \,, \\ \hat{\boldsymbol{v}}_\sigma^{(k)} &= \boldsymbol{v}_\sigma^{(k-1)}/(1-\rho_2^k) \quad \text{with} \quad \boldsymbol{v}_\sigma^{(k)} = \rho_2 \boldsymbol{v}_\sigma^{(k-1)} + (1-\rho_1)(\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)})^2 \,.\end{aligned} \tag{3.9.9}$$

The hyperparameters are set as follows: $\gamma = 10^{-3}, \rho_1 = 0.9, \rho_2 = 0.999, \epsilon = 10^{-8}$.

Chapter 4

Online Optimization of Non-convex Problems

Abstract: *Stochastic approximation (SA) is a key method used in statistical learning. Recently, its non-asymptotic convergence analysis has been considered in many papers. However, most of the prior analyses are made under restrictive assumptions such as unbiased gradient estimates and convex objective function, which significantly limit their applications to sophisticated tasks such as online and reinforcement learning. These restrictions are all essentially relaxed in this work. In particular, we analyze a general SA scheme to minimize a non-convex, smooth objective function. We consider update procedure whose drift term depends on a state-dependent Markov chain and the mean field is not necessarily of gradient type, covering approximate second-order method and allowing asymptotic bias for the one-step updates. We illustrate these settings with the online EM algorithm and the policy-gradient method for average reward maximization in reinforcement learning. This chapter corresponds to the article [Karimi et al., 2019a].*

Contents

4.1 Introduction

Stochastic Approximation (SA) schemes are sequential (online) methods for finding a zero of a function when only noisy observations of the function values are available. Consider the recursion:

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n - \gamma_{n+1} H_{\boldsymbol{\theta}_n}(X_{n+1}), \quad n \in \mathbb{N} \tag{4.1.1}$$

where $\boldsymbol{\theta}_n \in \Theta \subset \mathbb{R}^d$ denotes the nth iterate, $\gamma_n > 0$ is the step size and $H_{\boldsymbol{\theta}_n}(X_{n+1})$ is the nth *stochastic* update (a.k.a. drift term) depending on a random element X_{n+1} taking its values in a measurable space X. In the simplest setting, $\{X_n,\ n \in \mathbb{N}\}$ is an $\sim_{\text{i.i.d.}}$ sequence of random vectors and $H_{\boldsymbol{\theta}_n}(X_{n+1})$ is a conditionally *unbiased* estimate of the so-called mean-field $h(\boldsymbol{\theta}_n)$, *i.e.,* $\mathbb{E}\left[H_{\boldsymbol{\theta}_n}(X_{n+1}) \,|\, \mathcal{F}_n\right] = h(\boldsymbol{\theta}_n)$ where $\mathcal{F}_n$ denotes the filtration generated by the random variables $(\boldsymbol{\theta}_0, \{X_m\}_{m \le n})$. In such case, $\boldsymbol{e}_{n+1} = H_{\boldsymbol{\theta}_n}(X_{n+1}) - h(\boldsymbol{\theta}_n)$ is a *martingale difference*. In more sophisticated settings, $\{X_n,\ n \in \mathbb{N}\}$ is a *state-dependent* (or controlled) Markov chain, *i.e.,* for any bounded measurable function $f : \mathsf{X} \to \mathbb{R}$,

$$\mathbb{E}\left[f(X_{n+1}) \,|\, \mathcal{F}_n\right] = P_{\boldsymbol{\eta}_n} f(X_n) = \int f(x) P_{\boldsymbol{\eta}_n}(X_n, \mathrm{d}x) \,, \tag{4.1.2}$$

where $P_{\boldsymbol{\eta}} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$ is a Markov kernel such that, for each $\boldsymbol{\theta} \in \Theta$, $P_{\boldsymbol{\eta}}$ has a unique stationary distribution $\pi_{\boldsymbol{\theta}}$. In such case, the mean field for the SA is defined as:

$$h(\boldsymbol{\theta}) = \textstyle\int H_{\boldsymbol{\theta}}(x) \pi_{\boldsymbol{\theta}}(\mathrm{d}x) \,, \tag{4.1.3}$$

where we have assumed that $\int \|H_{\boldsymbol{\theta}}(x)\| \pi_{\boldsymbol{\theta}}(\mathrm{d}x) < \infty$.

Throughout this paper, we assume that the mean field h is 'related' (to be defined precisely later) to a smooth Lyapunov function $V : \mathbb{R}^d \to \mathbb{R}$, where $V(\boldsymbol{\theta}) > -\infty$. The aim of the SA scheme (4.1.1) is to find a minimizer or stationary point of the possibly non-convex Lyapunov function V.

Though more than 60 years old [Robbins and Monro, 1951], SA is now of renewed interest as it covers a wide range of applications at the heart of many successes with statistical learning. This includes in particular the stochastic gradient (SG) method and its variants as surveyed in [Bottou, 1998, Bottou et al., 2018], but also in reinforcement learning [Peters and Schaal, 2008, Sutton and Barto, 2018, Williams, 1992]. Most convergence analyses assume that $\{\boldsymbol{\theta}_n,\ n \in \mathbb{N}\}$ is bounded with probability one or visits a prescribed compact set infinitely often. Under such global stability or recurrence conditions [and appropriate regularity conditions on the mean field h], the SA sequences might be seen as approximation of the ordinary differential equation $\dot{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$. Most results available as of today [see for example [Benveniste et al., 1990], [Kushner and Yin, 2003, Chapter 5, Theorem 2.1] or [Borkar, 2009]] have an asymptotic flavor. The focus is to establish that the stationary point of the sequence $\{\boldsymbol{\theta}_n,\ n \in \mathbb{N}\}$ belongs to a stable attractor of its

limiting ODE.

To gain insights on the difference among statistical learning algorithms, non-asymptotic analysis of SA scheme has been considered only recently. In particular, SG methods whose mean field is the gradient of the objective function, *i.e.,* $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$, are considered by Moulines and Bach [2011] for strongly convex function V and martingale difference noise; see [Bottou et al., 2018] for a recent survey on the topic. Extensions to stationary dependent noise have been considered in [Agarwal and Duchi, 2013, Duchi et al., 2012]. Meanwhile, many machine learning models can lead to non-convex optimization problems. To this end, SG methods for non-convex, smooth objective function V have been first studied in [Ghadimi and Lan, 2013] with martingale noise (see [Bottou et al., 2018, Section 4]), and it was extended in [Sun et al., 2018] to the case where $\{X_n,\ n \in \mathbb{N}\}$ is a state-independent Markov chain, *i.e.,* the Markov kernel in (4.1.2) does not depend on $\boldsymbol{\theta}$.

Of course, SA schemes go far beyond SG methods. In fact, in many important applications, the drift term of the SA is *not* a noisy version of the gradient, *i.e.,* the mean field h is not the gradient of V. Obvious examples include second-order methods, which aim at combatting the adverse effects of high non-linearity and ill-conditioning of the objective function through stochastic quasi-Newton algorithms. Another closely related example is the online Expectation Maximization (EM) algorithm introduced by Cappé and Moulines [2009] and is further developed in [Balakrishnan et al., 2017, Chen et al., 2018]. In many cases, the mean field of the drift term may even be asymptotically biased with the random element $\{X_n,\ n \in \mathbb{N}\}$ drawn from a Markov chain with *state-dependent* transition probability. Examples for this situation are common in reinforcement learning such as Q-learning [Jaakkola et al., 1994], policy gradient [Baxter and Bartlett, 2001] and temporal difference learning [Bhandari et al., 2018, Dalal et al., 2018a,b, Lakshminarayanan and Szepesvari, 2018].

Surprisingly enough, we are not aware of non-asymptotic convergence results of the general SA (4.1.1) comparable to [Ghadimi and Lan, 2013] and [Bottou et al., 2018, Section 4,5] when (a) the drift term $H_{\boldsymbol{\theta}}(x)$ in (4.1.1) is not the noisy gradient of the objective function V and is potentially biased, and/or (b) the sequence $\{X_n,\ n \in \mathbb{N}\}$ is a *state-dependent* Markov chain. To this end, the main objective of this work is to fill this gap in the literature by establishing non-asymptotic convergence of SA under the above settings. Our main assumption is the existence of a smooth function V satisfying for all $\boldsymbol{\theta} \in \Theta$, $c_0 + c_1 \langle \nabla V(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle \geq \|h(\boldsymbol{\theta})\|^2$ there exists $c_1 > 0, c_0 \geq 0$; see Section 4.2 and H4.1. If $c_0 = 0$, then $\langle \nabla V(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle > 0$ as soon as $h(\boldsymbol{\theta}) \neq \mathbf{0}$ in which case V is a Lyapunov function for the ODE $\dot{\boldsymbol{\theta}} = h(\boldsymbol{\theta})$. Assuming $c_0 > 0$ allows us to consider situations in which the estimate of the mean field is biased, a situation which has been first studied in Tadić and Doucet [2017]. To summarize, our contributions are two-fold:

1. We provide *non-asymptotic* convergence analysis for (4.1.1) with a potentially biased

mean field h under two cases — (Case 1) $\{X_n,\ n \in \mathbb{N}\}$ is an $\sim_{\text{i.i.d.}}$ sequence; (Case 2) $\{X_n,\ n \in \mathbb{N}\}$ is a *state-dependent* Markov chain. For these two cases, we provide non asymptotic bounds such that for all $n \in \mathbb{N}$, $\mathbb{E}[\|h(\boldsymbol{\theta}_N)\|^2] = \mathcal{O}(c_0 + \log(n)/\sqrt{n})$, for some random index $N \in \{1, \ldots, n\}$ and $c_0 \geq 0$ characterizes the (potential) bias of the mean field h.

2. We illustrate our findings by analyzing popular statistical learning algorithms such as the online expectation maximization (EM) algorithm [Cappé and Moulines, 2009] and the average-cost policy-gradient method [Sutton and Barto, 2018]. Our findings provide new insights into the non-asymptotic convergence behavior of these algorithms.

Our theory significantly extends the results reported in [Bottou et al., 2018, Sections 4,5] and [Ghadimi and Lan, 2013, Theorem 2.1]. When focused on the Markov noise setting, our result is a nontrivial relaxation of [Sun et al., 2018], which considers Markov noise that is *not state dependent* and the mean field satisfies $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$; and of [Tadić and Doucet, 2017] which shows asymptotic convergence of (4.1.1) under the uniform boundedness assumption on iterates.

Notation Let $(\mathsf{X}, \mathcal{X})$ be a measurable space. A Markov kernel R on $\mathsf{X} \times \mathcal{X}$ is a mapping $R : \mathsf{X} \times \mathcal{X} \to [0, 1]$ satisfying the following conditions: (a) for every $x \in \mathsf{X}$, $R(x, \cdot) : A \mapsto R(x,\ A)$ is a probability measure on $\mathcal{X}$ (b) for every $\mathsf{A} \in \mathcal{X}$, $R(\cdot, A) : x \mapsto R(x, A)$ is a measurable function. For any probability measure λ on $(\mathsf{X}, \mathcal{X})$, we define λR by $\lambda R(\mathsf{A}) = \int_{\mathsf{X}} \lambda(\mathrm{d}x) R(x, \mathsf{A})$. For all $k \in \mathbb{N}^*$, we define the Markov kernel R^k recursively by $R^1 = R$ and for all $x \in \mathsf{X}$ and $\mathsf{A} \in \mathcal{X}$, $R^{k+1}(x, \mathsf{A}) = \int_{\mathsf{X}} R^k(x, \mathrm{d}x') R(x', \mathsf{A})$. A probability measure $\bar{\pi}$ is invariant for R if $\bar{\pi} R = \bar{\pi}$. $\|\cdot\|$ denotes the standard Euclidean norm (for vectors) or the operator norm (for matrices).

4.2 Stochastic Approximation Schemes and Their Convergence

Consider the following assumptions:

H4.1 *For all $\boldsymbol{\theta} \in \Theta$, there exists $c_0 \geq 0, c_1 > 0$ such that $c_0 + c_1 \langle \nabla V(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle \geq \|h(\boldsymbol{\theta})\|^2$.*

H4.2 *For all $\boldsymbol{\theta} \in \Theta$, there exists $d_0 \geq 0, d_1 > 0$ such that $d_0 + d_1 \|h(\boldsymbol{\theta})\| \geq \|\nabla V(\boldsymbol{\theta})\|$.*

H4.3 *Lyapunov function V is L-smooth. For all $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, $\|\nabla V(\boldsymbol{\theta}) - \nabla V(\boldsymbol{\theta}')\| \leq L\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.*

A4.1, H4.2 assume that the mean field $h(\boldsymbol{\theta})$ [cf. (4.1.2)] is indirectly related to the Lyapunov function $V(\boldsymbol{\theta})$ where it needs not be the same as $\nabla V(\boldsymbol{\theta})$. In particular, the constants c_0, d_0

characterize the 'bias' between the mean field and the gradient of the Lyapunov function. From an optimization perspective, we note that the Lyapunov function V can be *non-convex* under H4.3. In light of H4.1, H4.2, we study the convergence of the non-negative quantity $\|h(\boldsymbol{\theta}_n)\|^2$, where $\boldsymbol{\theta}_n$ is produced by (4.1.1). If $c_0 = d_0 = 0$ in H4.1, H4.2, then $h(\boldsymbol{\theta}_*) = 0$ implies that $\|\nabla V(\boldsymbol{\theta}_*)\| = 0$, *i.e.,* the point $\boldsymbol{\theta}_*$ is a stationary point of the deterministic recursion $\bar{\boldsymbol{\theta}}_n = \bar{\boldsymbol{\theta}}_n - \gamma_{n+1} h(\bar{\boldsymbol{\theta}}_n)$. As a convention, for any $\epsilon \geq 0$, we say that $\boldsymbol{\theta}_*$ is an *ϵ-stationary point* if $\|h(\boldsymbol{\theta}_*)\|^2 \leq \epsilon$.

As a common step in analyzing SA scheme for smooth but non-convex Lyapunov function (e.g., [Ghadimi and Lan, 2013]), we shall adopt a randomized stopping rule. For any $n \geq 1$, let $N \in \{0, \ldots, n\}$ be a discrete random variable (independent of $\{\mathcal{F}_n,\ n \in \mathbb{N}\}$) with

$$\mathbb{P}(N = \ell) := \left(\textstyle\sum_{k=0}^{n} \gamma_{k+1}\right)^{-1} \gamma_{\ell+1} \ , \tag{4.2.1}$$

where N serves as the terminating iteration for (4.1.1). Throughout this paper, we focus on analyzing $\mathbb{E}[\|\nabla h(\boldsymbol{\theta}_N)\|^2]$ where the expectation is taken over N and the stochastic updates in SA. We consider two settings for the noise in SA scheme. Define the following noise vector:

$$\boldsymbol{e}_{n+1} := H_{\boldsymbol{\theta}_n}(X_{n+1}) - h(\boldsymbol{\theta}_n) \ , \tag{4.2.2}$$

where $h(\boldsymbol{\theta}_n)$ was defined in (7.2.17). Our settings and convergence results are in order.

Case 1. $\{\boldsymbol{e}_n\}_{n\geq 1}$ **is a Martingale Difference Sequence.** We first consider a case similar to the classical SG method analyzed by Ghadimi and Lan [2013]. In particular,

H4.4 *The sequence of noise vectors is a Martingale difference sequence with, for any* $n \in \mathbb{N}$, $\mathbb{E}\left[\boldsymbol{e}_{n+1} \,|\, \mathcal{F}_n\right] = \mathbf{0}$, $\mathbb{E}\left[\|\boldsymbol{e}_{n+1}\|^2 \,|\, \mathcal{F}_n\right] \leq \sigma_0^2 + \sigma_1^2 \|h(\boldsymbol{\theta}_n)\|^2$ *with* $\sigma_0^2, \sigma_1^2 \in [0, \infty)$.

As a concrete example, H4.4 can be satisfied when $H_{\boldsymbol{\theta}_n}(X_{n+1}) = h(\boldsymbol{\theta}_n) + X_{n+1}$ where X_{n+1} is an i.i.d., zero-mean random vector with bounded variance. We show:

Theorem 3 *Let H4.1, H4.3, H4.4 hold and* $\gamma_{n+1} \leq (2c_1 L(1+\sigma_1^2))^{-1}$ *for all* $n \geq 0$. *We have*

$$\mathbb{E}[\|h(\boldsymbol{\theta}_N)\|^2] \leq \frac{2c_1(V_{0,n} + \sigma_0^2 L \sum_{k=0}^{n} \gamma_{k+1}^2)}{\sum_{k=0}^{n} \gamma_{k+1}} + 2c_0 \ , \tag{4.2.3}$$

where N *is distributed according to* (4.2.1) *and we have defined* $V_{0,n} := \mathbb{E}[V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}_{n+1})]$.

Proof The proof is postponed to Section 4.2.1

If we set $\gamma_k = (2c_1 L(1+\sigma_1^2)\sqrt{k})^{-1}$ for all $k \geq 1$, then the right hand side in (4.2.3) evaluates to $\mathcal{O}(c_0 + \log n/\sqrt{n})$ for any $n \geq 1$. Therefore, the SA scheme (4.1.1) finds an $\mathcal{O}(c_0 + \log n/\sqrt{n})$ stationary point within n iterations.

Case 2. $\{\boldsymbol{e}_n\}_{n\geq 1}$ **is State-dependent Markov Noise.** Next, we consider a general scenario when X_{n+1} is drawn from a state-dependent Markov process, *i.e.,* for any bounded measurable function φ and $n \in \mathbb{N}$, we have $\mathbb{E}\left[\varphi(X_{n+1}) \,|\, \mathcal{F}_n\right] = P_{\boldsymbol{\eta}_n}\varphi(X_n)$, where $P_{\boldsymbol{\eta}}$ is a Markov kernel on $\mathsf{X} \times \mathcal{X}$. We assume that for each $\boldsymbol{\theta} \in \Theta$, $P_{\boldsymbol{\eta}}$ has a unique stationary distribution $\pi_{\boldsymbol{\theta}}$, *i.e.,* $\pi_{\boldsymbol{\theta}} P_{\boldsymbol{\eta}} = \pi_{\boldsymbol{\theta}}$. In addition, for each $\boldsymbol{\theta} \in \Theta$, we have $\int \|H_{\boldsymbol{\theta}}(x)\| \pi_{\boldsymbol{\theta}}(\mathrm{d}x) < \infty$ and $h(\boldsymbol{\theta}) = \int H_{\boldsymbol{\theta}}(x)\pi_{\boldsymbol{\theta}}(\mathrm{d}x)$. Consider the following assumptions that are similar to [Tadić and Doucet, 2017, Section 3]:

H4.5 *There exists a Borel measurable function* $\hat{H} : \boldsymbol{\theta} \times \mathsf{X} \to \boldsymbol{\theta}$ *where for each* $\boldsymbol{\theta} \in \Theta$, $x \in \mathsf{X}$,

$$\hat{H}_{\boldsymbol{\theta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\theta}}(x) = H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})\,. \tag{4.2.4}$$

H4.6 *There exists* $L_{PH}^{(0)} < \infty$ *and* $L_{PH}^{(1)} < \infty$ *such that, for all* $\boldsymbol{\theta} \in \Theta$ *and* $x \in \mathsf{X}$, *one has* $\|\hat{H}_{\boldsymbol{\theta}}(x)\| \leq L_{PH}^{(0)}$, $\|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\theta}}(x)\| \leq L_{PH}^{(0)}$. *Moreover, for* $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$,

$$\sup_{x\in\mathsf{X}} \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\theta}}(x) - P_{\boldsymbol{\theta}'}\hat{H}_{\boldsymbol{\theta}'}(x)\| \leq L_{PH}^{(1)}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|\,. \tag{4.2.5}$$

H4.7 *The stochastic update is bounded,* i.e., $\sup_{\boldsymbol{\theta}\in\Theta, x\in\mathsf{X}} \|H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})\| \leq \sigma$.

Basically, assumption H4.5 requires that for each $\boldsymbol{\theta} \in \Theta$, the Poisson equation associated with the Markov kernel $P_{\boldsymbol{\eta}}$ and the function $H_{\boldsymbol{\theta}}(\cdot)$ has a solution. Assumption H4.6 implies that for each $x \in \mathsf{X}$, the function $\boldsymbol{\theta} \mapsto H_{\boldsymbol{\theta}}(x)$ is Lipshitz and that the Lipshitz constant is uniformly bounded in $x \in \mathsf{X}$. We provide in Appendix 4.8 conditions upon which these assumptions hold. Lastly, Assumption H4.7 assumes that the drift terms are bounded uniformly. Our main result reads as follows:

Theorem 4 *Let H4.1–H4.3, H4.5–H4.7 hold. Suppose that the step sizes satisfy*

$$\gamma_{n+1} \leq \gamma_n,\ \gamma_n \leq a\gamma_{n+1},\ \gamma_n - \gamma_{n+1} \leq a'\gamma_n^2,\ \gamma_1 \leq 0.5\big(c_1(L + C_h)\big)^{-1}\,, \tag{4.2.6}$$

for some $a, a' > 0$ *and all* $n \geq 0$. *We have*

$$\mathbb{E}[h(\boldsymbol{\theta}_N)\|^2] \leq \frac{2c_1\big(V_{0,n} + C_{0,n} + (\sigma^2 L + C_\gamma)\sum_{k=0}^n \gamma_{k+1}^2\big)}{\sum_{k=0}^n \gamma_{k+1}} + 2c_0\,, \tag{4.2.7}$$

where N *is distributed according to* (4.2.1), $V_{0,n} := \mathbb{E}[V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}_{n+1})]$, *and the constants are:*

$$C_h := \big(L_{PH}^{(1)}(d_0 + \frac{d_1}{2}(a+1) + ad_1\sigma) + L_{PH}^{(0)}(L + d_1\{1 + a'\})\big)\,, \tag{4.2.8}$$

$$C_\gamma := L_{PH}^{(1)}(d_0 + d_0\sigma + d_1\sigma) + LL_{PH}^{(0)}(1+\sigma)\,, \tag{4.2.9}$$

$$C_{0,n} := L_{PH}^{(0)}\big((1+d_0)(\gamma_1 - \gamma_{n+1}) + d_0(\gamma_1 + \gamma_{n+1}) + 2d_1\big)\,. \tag{4.2.10}$$

Proof The proof is postponed to Section 4.2.1

Similar to the case with Martingale difference noise, if we set $\gamma_k = (2c_1L(1+C_h)\sqrt{k})^{-1}$ for all $k \geq 1$, then the step size satisfies (4.2.6) with $a = \sqrt{2}$ and $a' = \frac{\sqrt{2}-1}{\sqrt{2}}(2c_1L(1+C_h))$,

and the right hand side in (4.2.7) evaluates to $\mathcal{O}(c_0 + \log n/\sqrt{n})$ for any $n \geq 1$. We obtain a similar convergence rate as in Theorem 3. In fact, if we consider a special case when for all $\boldsymbol{\theta} \in \Theta$ and $x \in \mathsf{X}$, $P_{\boldsymbol{\eta}}(x, \cdot) = \pi_{\boldsymbol{\theta}}(\cdot)$, we have $L_{PH}^{(0)} = L_{PH}^{(1)} = 0$. The constants evaluates to $C_h = C_\gamma = C_{0,n} = 0$ and our Theorem 4 can be reduced into Theorem 3. We remark that Theorem 4 cannot be treated as a strict generalization of Theorem 3 as H4.4 does not imply the uniform boundedness H4.7.

Our analysis [cf. Lemma 4] relies on a new decomposition of the error terms. This allows us to control the growth of $\mathbb{E}[\|h(\boldsymbol{\theta}_n)\|^2]$ with $\boldsymbol{\theta}_n$ produced by the SA scheme without explicitly assuming that $\{\boldsymbol{\theta}_n\}_{n\geq 0}$ is bounded.

Lower Bound We provide a lower bound on $\mathbb{E}[\|h(\boldsymbol{\theta}_N)\|^2]$ with the SA scheme (4.1.1) and (4.2.1):

Lemma 2 *Consider the SA scheme (4.1.1) with $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$. There exists a Lyapunov function $V(\boldsymbol{\theta})$ satisfying H4.3 and a noise sequence $\{\boldsymbol{e}_n\}_{n\geq 1}$ satisfying H4.4-H4.7 such that for any $n \geq 1$,*

$$\mathbb{E}[\|h(\boldsymbol{\theta}_N)\|^2] \geq \frac{\mathbb{E}[V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}_{n+1})] + C_{\mathsf{lb}} \sum_{k=0}^n \gamma_{k+1}^2}{\sum_{k=0}^n \gamma_{k+1}} \tag{4.2.11}$$

where N is distributed according to (4.2.1), and $C_{\mathsf{lb}} > 0$ is some constant independent of n.

Proof The proof is postponed to Appendix 4.5.3

For large n, setting $\gamma_k = c/\sqrt{k}$ minimizes the right hand side of (4.2.11), yielding $\mathbb{E}[\|h(\boldsymbol{\theta}_N)\|^2] = \Omega(\log(n)/\sqrt{n})$. The considered SA scheme satisfies assumptions H4.1-H4.7, and the lower bound (4.2.11) matches the upper bounds in Theorem 3 & 4 (when $c_0 = 0$). The upper bounds are therefore tight.

We remark that our proof in Appendix 4.5.3 uses the construction with a strongly convex Lyapunov function. It does not violate the known $\mathbb{E}[\|h(\frac{1}{n+1}\sum_{k=0}^n \boldsymbol{\theta}_k)\|^2] = \mathcal{O}(1/n)$ rate in [Moulines and Bach, 2011] as the latter uses SA with a Polyak-Ruppert average estimator. To our best knowledge, it remains an open problem to lower bound the convergence rate of SA for smooth but non-convex Lyapunov function. We mention here a recent work [Fang et al., 2018, Remark 1] which shows $\mathbb{E}[\|h(\boldsymbol{\theta}_n)\|^2] = \Omega(1/\sqrt{n})$ under different conditions than those satisfied in this paper.

Related Studies Non-asymptotic analysis of biased SA schemes can be found in the literature on temporal difference (TD) learning [Bhandari et al., 2018, Dalal et al., 2018a,b, Lakshminarayanan and Szepesvari, 2018], which analyzed a special case of linear SA. Their assumptions can essentially be covered by our H4.1–H4.3 with $V(\boldsymbol{\theta}) = \|\boldsymbol{\theta} - \boldsymbol{\theta}^\star\|_\Phi^2$, e.g., [Bhandari et al., 2018, Lemma 3] shows that the TD learning has a mean field which

satisfies H4.1. Furthermore, we note that the above mentioned analysis are based on a strongly convex Lyapunov function.

For Case 1, our results generalizes [Ghadimi and Lan, 2013, Theorem 2.1] by accounting for biased SA updates. In fact we recover the latter result with $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$, H4.1 $[c_0 = 0, c_1 = 1]$.

For Case 2, our assumptions H4.1–H4.3, H4.5–H4.7 are similar to [Tadić and Doucet, 2017, Section 3]. The exception is H4.7 which is used in place of the implicit assumption $\sup_{n\in\mathbb{N}} \|\boldsymbol{\theta}_n\| < \infty$ of the latter. However, we note that the two conditions are neither stronger nor weaker than the other.

4.2.1 Convergence Analysis

The detailed proofs in this section are in Appendix 4.5. To simplify notations, we shall denote $h_n := \|h(\boldsymbol{\theta}_n)\|^2$ from now on. We first describe an intermediate result that holds under just H4.1, H4.3:

Lemma 3 *Let H4.1, H4.3 hold. It holds for all $n \geq 1$ that:*

$$\begin{aligned}&\textstyle\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}(1 - c_1 L\gamma_{k+1})h_k \\ &\leq V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}_{n+1}) + L\textstyle\sum_{k=0}^{n} \gamma_{k+1}^2 \|\boldsymbol{e}_{k+1}\|^2 + \sum_{k=0}^{n} \gamma_{k+1}(\frac{c_0}{c_1} - \langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1}\rangle\,).\end{aligned} \tag{4.2.12}$$

Proof The proof is postponed to Appendix 4.5.1

Proof of Theorem 3 Having established Lemma 3, the convergence of SA with Martingale difference noise can be obtained. Particularly, the expected value of $\langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1}\rangle$ is zero when conditioned on $\mathcal{F}_k$. Therefore, taking total expectation on both sides of (4.2.12) yields:

$$\begin{aligned}\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}(1 - c_1 L\gamma_{k+1})\mathbb{E}[h_k] &\leq V_{0,n} + L\sum_{k=0}^{n} (\gamma_{k+1}^2 \mathbb{E}[\|\boldsymbol{e}_{k+1}\|^2] + \gamma_{k+1}\frac{c_0}{c_1}) \\ &\leq V_{0,n} + L\sigma_0^2 \sum_{k=0}^{n} \gamma_{k+1}^2 + L\sigma_1^2 \sum_{k=0}^{n} \gamma_{k+1}\mathbb{E}[h_k]) + \gamma_{k+1}\frac{c_0}{c_1}) \,,\end{aligned} \tag{4.2.13}$$

where the last inequality is due to H4.4. Rearranging terms yields:

$$\textstyle\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}(1 - c_1 L(1+\sigma_1^2)\gamma_{k+1})\mathbb{E}[h_k] \leq V_{0,n} + \sigma_0^2 L\sum_{k=0}^{n} \gamma_{k+1}^2 + \frac{c_0}{c_1}\sum_{k=0}^{n} \gamma_{k+1} \,. \tag{4.2.14}$$

Consequently, using (4.2.1) and noting that $1 - c_1 L(1+\sigma_1^2)\gamma_{n+1} \geq \frac{1}{2}$, we obtain

$$\mathbb{E}[h_N] = \sum_{n'=0}^{n} \frac{\gamma_{n'+1}\mathbb{E}[h_{n'}]}{\sum_{k=0}^{n}\gamma_{k+1}} \leq \frac{2c_1\left(V_{0,n} + \sigma_0^2 L \sum_{k=0}^{n}\gamma_{k+1}^2\right)}{\sum_{k=0}^{n}\gamma_{k+1}} + 2c_0 \ . \tag{4.2.15}$$

Proof of Theorem 4 In the case with state-dependent Markovian noise. Under H4.7, one has

$$\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[\|\boldsymbol{e}_{k+1}\|^2] \leq \sum_{k=0}^{n}\gamma_{k+1}^2\sigma^2 \ . \tag{4.2.16}$$

Unlike in Theorem 3, the expected value of the inner product $\langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1}\rangle$ is non-zero in general. Fortunately, as we show next in Lemma 4, this issue can be mitigated.

Lemma 4 *Let H4.1–H4.3, H4.5–H4.7 hold and the step sizes satisfy (4.2.6). It holds:*

$$\mathbb{E}\Big[-\sum_{k=0}^{n}\gamma_{k+1}\langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1}\rangle\Big] \leq C_h\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[\|h(\boldsymbol{\theta}_k)\|^2] + C_\gamma\sum_{k=0}^{n}\gamma_{k+1}^2 + C_{0,n} \ , \tag{4.2.17}$$

where C_h, C_γ and $C_{0,n}$ are defined in (4.2.8), (4.2.9), (4.2.10).

Proof The proof is postponed to Appendix 4.5.2

Finally, to prove the theorem, we combine Lemma 3, (4.2.16) and Lemma 4 to obtain:

$$\begin{aligned} &\textstyle\sum_{k=0}^{n}\frac{\gamma_{k+1}}{c_1}(1 - c_1 L\gamma_{k+1})\mathbb{E}[h_k] \\ &\leq V_{0,n} + C_{0,n} + \left(\sigma^2 L + C_\gamma\right)\textstyle\sum_{k=0}^{n}\gamma_{k+1}^2 + C_h\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[h_k] + \frac{c_0}{c_1}\sum_{k=0}^{n}\gamma_{k+1} \ . \end{aligned} \tag{4.2.18}$$

Repeating a similar argument as in (4.2.15) using the distribution (4.2.1) shows the desired bound (4.2.7).

4.3 Application to Online and Reinforcement Learning

In this section, we present several applications pertaining to machine learning where the results in Section 4.2 apply and provide new non-asymptotic convergence rate for them.

4.3.1 Regularized Online Expectation Maximization

Expectation-Maximization (EM) [Dempster et al., 1977] is a powerful tool for learning latent variable models, which can be inefficient due to the high storage cost. This has motivated the development of online version of the EM which makes it possible to estimate the parameters of latent variables model without storing the data; the online EM algorithm analyzed below was introduced in [Cappé and Moulines, 2009] and later developed by many authors: see for example [Chen et al., 2018] and the references therein. The online EM

algorithm sticks closely to the principles of the batch-mode EM algorithm. Each iteration of the online EM algorithm is decomposed into two steps, where the first one is a stochastic approximation version of the E-step aimed at incorporating the information brought by the newly available observation, and, the second step consists in the maximization program that appears in the M-step of the traditional EM algorithm.

The latent variable statistical model postulates the existence of a latent variable X distributed under $f(x;\boldsymbol{\theta})$ where $\{f(x;\boldsymbol{\theta});\boldsymbol{\theta}\in\Theta\}$ is a parametric family of probability density functions and Θ is an open convex subset of $\mathbb{R}^d$. The observation $Y\in\mathsf{Y}$ is a deterministic function of X. We denote by $g(y;\boldsymbol{\theta})$ the (observed) likelihood function. The notations $\mathbb{E}_{\boldsymbol{\theta}}[\cdot]$ and $\mathbb{E}_{\boldsymbol{\theta}}[\cdot\,|\,Y]$ are used to denote the expectation and conditional expectation under the statistical model $\{f(x;\boldsymbol{\theta});\boldsymbol{\theta}\in\Theta\}$. We denote by π the probability density function of the observation Y: the model might be misspecified, that is, the "true" distribution of the observations may not belong to the family $\{g(y;\boldsymbol{\theta}),\boldsymbol{\theta}\in\Theta\}$. The notations $\mathbb{E}_{\pi}$ is used below to denote the expectation under the actual distribution of the observations. Let S be a convex open subset of $\mathbb{R}^m$ and $S:\mathsf{X}\to\mathsf{S}$ be a measurable function. We assume that the complete data-likelihood function belongs to the curved exponential family

$$f(x;\boldsymbol{\theta})=h(x)\exp\left(\langle S(x)\,|\,\phi(\boldsymbol{\theta})\rangle-\psi(\boldsymbol{\theta})\right)\ , \tag{4.3.1}$$

where $\psi:\Theta\to\mathbb{R}$ is twice differentiable and convex and $\phi:\Theta\to\mathsf{S}\subset\mathbb{R}^m$ is concave and differentiable. In this setting, S is the complete data sufficient statistics. For any $\boldsymbol{\theta}\in\Theta$ and $y\in\mathsf{Y}$, we assume that the conditional expectation

$$\overline{\boldsymbol{s}}(y;\boldsymbol{\theta})=\mathbb{E}_{\boldsymbol{\theta}}\left[S(X)\,|\,Y=y\right] \tag{4.3.2}$$

is well-defined and belongs to S. For any $\boldsymbol{s}\in\mathsf{S}$, we consider the penalized negated complete data log-likelihood defined as

$$\ell(\boldsymbol{s};\boldsymbol{\theta}):=\psi(\boldsymbol{\theta})+\mathrm{R}(\boldsymbol{\theta})-\langle\boldsymbol{s}\,|\,\phi(\boldsymbol{\theta})\rangle\ , \tag{4.3.3}$$

where $\mathrm{R}:\Theta\mapsto\mathbb{R}$ is a penalization term assumed to be twice differentiable. This penalty term is used to enforce constraints on the estimated parameter. If $\kappa:\Theta\to\mathbb{R}^m$ is a differentiable function, we denote by $\mathrm{J}_{\kappa}^{\boldsymbol{\theta}}(\boldsymbol{\theta}')\in\mathbb{R}^{m\times d}$ the Jacobian of the map κ with respect to $\boldsymbol{\theta}$ at $\boldsymbol{\theta}'$. Consider:

H4.8 *For all $\boldsymbol{s}\in\mathsf{S}$, the function $\boldsymbol{\theta}\mapsto\ell(\boldsymbol{s};\boldsymbol{\theta})$ admits a unique global minimum in the interior of Θ, denoted by $\overline{\boldsymbol{\theta}}(\boldsymbol{s})$ and characterized by*

$$\nabla\psi(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))+\nabla\,\mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))-\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top}\boldsymbol{s}=\mathbf{0}\ . \tag{4.3.4}$$

In addition, for any $\boldsymbol{s}\in\mathsf{S}$, $\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$ is invertible and the map $\boldsymbol{s}\mapsto\overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is differentiable on S.

The *regularized* version of the online EM (ro-EM) method is an iterative procedure which alternatively updates an estimate of the sufficient statistics and the estimated parameters as:

$$\hat{\boldsymbol{s}}_{n+1} = \hat{\boldsymbol{s}}_n + \gamma_{n+1}(\overline{\boldsymbol{s}}(Y_{n+1}; \hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{s}}_n), \quad \hat{\boldsymbol{\theta}}_{n+1} = \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_{n+1}) \; . \tag{4.3.5}$$

In the following, we show that our *non-asymptotic* convergence result holds for the ro-EM. We establish convergence of the online method to a stationary point of the Lyapunov function defined as a regularized Kullback-Leibler (KL) divergence between π and $g_{\boldsymbol{\theta}}$. Precisely, we set

$$V(\boldsymbol{s}) := \mathrm{KL}\left(\pi, g(\cdot; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\right) + \mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) \quad \mathrm{KL}\left(\pi, g(\cdot; \boldsymbol{\theta})\right) := \mathbb{E}_\pi\left[\log(\pi(Y))/g(Y;\theta)\right] \; . \tag{4.3.6}$$

We establish a few key results that relate the ro-EM method to an SA scheme seeking for a stationary point of $V(\boldsymbol{s})$. Denote by $\mathcal{F}_n$ the filtration generated by the random variables $\{\hat{\boldsymbol{s}}_0, Y_k\}_{k \le n}$. From (5.2.5) we can identify the drift term and its mean field respectively as

$$\begin{aligned} &H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) = \hat{\boldsymbol{s}}_n - \overline{\boldsymbol{s}}(Y_{n+1}; \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_n)) \; , \\ &h(\hat{\boldsymbol{s}}_n) = \mathbb{E}_\pi\left[H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) | \mathcal{F}_n\right] = \hat{\boldsymbol{s}}_n - \mathbb{E}_\pi\left[\overline{\boldsymbol{s}}(Y_{n+1}; \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}_n))\right] \; . \end{aligned} \tag{4.3.7}$$

and $\boldsymbol{e}_{n+1} := H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) - h(\hat{\boldsymbol{s}}_n)$. Define by $\mathrm{H}_\ell^{\boldsymbol{\theta}}$ the Hessian of the function ℓ with respect to $\boldsymbol{\theta}$. Our results are summarized by the following propositions:

Proposition 5 *Assume H5.4. Then*

- *If $h(\boldsymbol{s}^\star) = \boldsymbol{0}$ for some $\boldsymbol{s}^\star \in \mathsf{S}$, then $\nabla_{\boldsymbol{\theta}} \mathrm{KL}\left(\pi, g_{\boldsymbol{\theta}^\star}\right) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\boldsymbol{\theta}^\star) = \boldsymbol{0}$ with $\boldsymbol{\theta}^\star := \overline{\boldsymbol{\theta}}(\boldsymbol{s}^\star)$.*
- *If $\nabla_{\boldsymbol{\theta}} \mathrm{KL}\left(\pi, g_{\boldsymbol{\theta}^\star}\right) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\boldsymbol{\theta}^\star) = \boldsymbol{0}$ for some $\boldsymbol{\theta}^\star \in \Theta$ then $\boldsymbol{s}^\star = \mathbb{E}_\pi[S(Y, \boldsymbol{\theta}^\star)]$.*

Proposition 6 *Assume H5.4. Then, for $\boldsymbol{s} \in \mathsf{S}$,*

$$\nabla_{\boldsymbol{s}} V(\boldsymbol{s}) = \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s}; \boldsymbol{\theta})\Big)^{-1} \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top \, h(\boldsymbol{s}) \; . \tag{4.3.8}$$

Proof The proofs are postponed to Appendix 4.6

Proposition 5 relates the root(s) of the mean field $h(\boldsymbol{s})$ to the stationary condition of the regularized KL divergence. Together with an additional condition on the smallest eigenvalue of the Jacobian-Hessian-Jacobian product

$$\lambda_{\mathsf{min}}\big(\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s}; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1} \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top\big) \ge \upsilon > 0, \; \forall \; \boldsymbol{s} \in \mathsf{S} \; , \tag{4.3.9}$$

Proposition 6 shows that the mean field of the stochastic update in (4.3.7) satisfies H4.1 with $c_0 = 0$ and $c_1 = 1/\upsilon$. If we assume that the Lyapunov function in (4.3.6), and the stochastic update in (4.3.7) satisfy the assumptions in Case 1 [*i.e.,* H4.4], then these

results show that within n iterations, the ro-EM method finds an $\mathcal{O}(\log n/\sqrt{n})$ stationary solution of the Lyapunov function. To further illustrate the above principles, we look at an example with Gaussian mixture model (GMM).

Example: GMM Inference Consider the inference problem of a mixture of M Gaussian distributions, each with a unit variance from an observation stream $Y_1, Y_2, \ldots$. The likelihood is:

$$g(y;\boldsymbol{\theta}) \propto \Big(1 - \sum_{m=1}^{M-1} \omega_m\Big) \exp\left(-\frac{(y-\mu_M)^2}{2}\right) + \sum_{m=1}^{M-1} \omega_m \exp\left(-\frac{(y-\mu_m)^2}{2}\right) . \quad (4.3.10)$$

The parameters are denoted by $\boldsymbol{\theta} := (\omega_1, \ldots, \omega_{M-1}, \mu_1, \ldots, \mu_{M-1}, \mu_M) \in \mathcal{C}$ where the parameter set is defined as $\mathcal{C} = \Delta_{M-1} \times \mathbb{R}^M$ with $\Delta_{M-1} := \{(\omega_1, \cdots, \omega_{M-1}) \in \mathbb{R}^{M-1}, \omega_m \geq 0, \sum_{m=1}^{M-1} \omega_m \leq 1\}$. To apply the ro-EM method, we augment the nth data Y_n with the latent variable $Z_n \in \{1, \ldots, M\}$. The log likelihood of the complete data tuple is

$$\mathcal{L}(\boldsymbol{x};\boldsymbol{\theta}) = \mathbb{1}_{\{z=M\}} \left[\log(1 - \sum_{m=1}^{M-1} \omega_m) - \frac{(y-\mu_M)^2}{2}\right] + \sum_{m=1}^{M-1} \mathbb{1}_{\{z=m\}} \left[\log(\omega_m) - \frac{(y-\mu_m)^2}{2}\right] . \quad (4.3.11)$$

The above can be written in the standard curved exponential family form (4.3.1). In particular, we partition the sufficient statistics as $S(\boldsymbol{x}) = (S^{(1)}(\boldsymbol{x})^\top, S^{(2)}(\boldsymbol{x})^\top, S^{(3)}(\boldsymbol{x}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, and partition $\phi(\boldsymbol{\theta}) = (\phi^{(1)}(\boldsymbol{\theta})^\top, \phi^{(2)}(\boldsymbol{\theta})^\top, \phi^{(3)}(\boldsymbol{\theta}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$. Using the fact that $\mathbb{1}_{\{z=M\}} = 1 - \sum_{m=1}^{M-1} \mathbb{1}_{\{z=m\}}$, (7.3.1) can be expressed in the standard form as (4.3.1) with

$$\begin{aligned} &s_m^{(1)} = \mathbb{1}_{\{z=m\}}, \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{\log(\omega_m) - \frac{\mu_m^2}{2}\right\} - \left\{\log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2}\right\} , \\ &s_m^{(2)} = \mathbb{1}_{\{z=m\}} y, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m, \quad m = 1, \ldots, M-1, \quad s^{(3)} = y, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M , \end{aligned} \quad (4.3.12)$$

and $\psi(\boldsymbol{\theta}) = -\left\{\log(1 - \sum_{j=1}^{M-1} \omega_j) - \frac{\mu_M^2}{2\sigma^2}\right\}$.

We apply the ro-EM method to the above model. Following the partition of sufficient statistics and parameters in the above, we define $\hat{\boldsymbol{s}}_n = ((\hat{\boldsymbol{s}}_n^{(1)})^\top, (\hat{\boldsymbol{s}}_n^{(2)})^\top, \hat{s}^{(3)})^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, and $\hat{\boldsymbol{\theta}}_n = (\hat{\boldsymbol{\omega}}_n^\top, \hat{\boldsymbol{\mu}}_n^\top, \hat{\mu}_M)^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$. Also, define the conditional expected value:

$$\widetilde{\omega}_m(Y_{n+1}; \hat{\boldsymbol{\theta}}_n) := \mathbb{E}_{\hat{\boldsymbol{\theta}}_n}[\mathbb{1}_{\{z=m\}} | Y = Y_{n+1}] = \frac{\hat{\omega}_{m,n} \exp(-\frac{1}{2}(Y_{n+1} - \hat{\mu}_{m,n})^2)}{\sum_{j=1}^{M} \hat{\omega}_{j,n} \exp(-\frac{1}{2}(Y_{n+1} - \hat{\mu}_{j,n})^2)} . \quad (4.3.13)$$

With the above notations, the E-step's update in (4.3.2) can be described with

$$\overline{s}(Y_{n+1};\hat{\boldsymbol{\theta}}_n) = \begin{pmatrix} \big(\widetilde{\omega}_1(Y_{n+1};\hat{\boldsymbol{\theta}}_n),\ldots,\widetilde{\omega}_{M-1}(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big)^\top \\ \big(Y_{n+1}\widetilde{\omega}_1(Y_{n+1};\hat{\boldsymbol{\theta}}_n),\ldots,Y_{n+1}\widetilde{\omega}_{M-1}(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big)^\top \\ Y_{n+1} \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{s}}_n^{(1)} \\ \overline{\boldsymbol{s}}_n^{(2)} \\ \overline{s}_n^{(3)} \end{pmatrix} . \tag{4.3.14}$$

For the M-step, let $\epsilon > 0$ be a user designed parameter, we consider the following regularizer:

$$\mathrm{R}(\boldsymbol{\theta}) = \epsilon \textstyle\sum_{m=1}^{M} \{\mu_m^2/2 - \log(\omega_m)\} - \epsilon \log\left(1 - \sum_{m=1}^{M-1} \omega_m\right) , \tag{4.3.15}$$

For any $\boldsymbol{s}$ with $\boldsymbol{s}^{(1)} \geq \mathbf{0}$, it can be shown that the regularized M-step in (5.2.5) evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1+\epsilon M)^{-1}\big(s_1^{(1)}+\epsilon,\ldots,s_{M-1}^{(1)}+\epsilon\big)^\top \\ \big((s_1^{(1)}+\epsilon)^{-1}s_1^{(2)},\ldots,(s_{M-1}^{(1)}+\epsilon)^{-1}s_{M-1}^{(2)}\big)^\top \\ \big(1-\sum_{m=1}^{M-1} s_m^{(1)}+\epsilon\big)^{-1}\big(s^{(3)}-\sum_{m=1}^{M-1} s_m^{(2)}\big) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{4.3.16}$$

Note that, as opposed to an unregularized solution (*i.e.,* with $\epsilon = 0$), the regularized solution is numerically stable as it avoids issues such as division by zero.

To analyze the convergence of ro-EM, we verify that (5.2.5), (4.3.14), (7.3.7) yield a special case of an SA scheme on $\hat{\boldsymbol{s}}_n$ which satisfies H4.1, H4.3, H4.4. Assume the following on the observations $\{Y_n\}_{n\geq 0}$

H4.9 *Each observed sample Y_n is drawn i.i.d. and they are bounded as $|Y_n| \leq \overline{Y}$ for any $n \geq 0$.*

The ro-EM method is initialized by setting $\hat{\boldsymbol{s}}_1 = (\mathbf{0},\mathbf{0},0)^\top$ and begun with the M-step. Note that under H4.9, the sufficient statistics $\hat{\boldsymbol{s}}_n$ lie in the compact set $\mathsf{S} = \Delta_{M-1} \times [-\overline{Y},\overline{Y}]^M$ for all $n \geq 1$, where $\Delta_{M-1} := \{s_1,\ldots,s_{M-1} : s_m \geq 0,\ \sum_{m=1}^{M-1} s_m \leq 1\}$. We observe the following propositions:

Proposition 7 *Under H4.9, it holds that $\mathbb{E}[\|\overline{s}(Y_{n+1};\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{s}}_n\|^2|\mathcal{F}_n] \leq 2M\overline{Y}^2$ for all $n \geq 0$.*

Proposition 8 *Under H4.9 and the regularizer (7.3.6) set with $\epsilon > 0$, then for all $(\boldsymbol{s},\boldsymbol{s}') \in \mathsf{S}^2$, there exists positive constants υ, Υ, Ψ such that:*

$$\langle \nabla V(\boldsymbol{s}) \,|\, h(\boldsymbol{s})\rangle \geq \upsilon \,\|h(\boldsymbol{s})\|^2, \quad \|\nabla V(\boldsymbol{s}) - \nabla V(\boldsymbol{s}')\| \leq \Psi\|\boldsymbol{s}-\boldsymbol{s}'\| . \tag{4.3.17}$$

Proof The proofs are postponed to Appendix 4.6

The above propositions show that the ro-EM method applied to GMM is a special case of the SA scheme with Martingale difference noise, for which H4.1 [with $c_0 = 0$, $c_1 = \upsilon^{-1}$],

and H4.3 [with $L = \Psi$], H4.4 [with $\sigma_0^2 = 2M\overline{Y}^2$, $\sigma_1^2 = 0$] are satisfied. As such, applying Theorem 3 shows that

Corollary 1 *Under H4.9 and set* $\gamma_k = (2c_1L(1+\sigma_1^2)\sqrt{k})^{-1}$. *For any* $n \in \mathbb{N}$, *let* $N \in \{0, ..., n\}$ *be an independent discrete r.v. distributed according to* (4.2.1). *The ro-EM method for GMM* (5.2.5), (4.3.14), (7.3.7) *finds a sufficient statistics such that*

$$\mathbb{E}[\|\nabla V(\hat{\boldsymbol{s}}_N)\|^2] = \mathcal{O}(\log n/\sqrt{n}) \tag{4.3.18}$$

where $V(\cdot)$ *is defined in* (4.3.6). *The expectation above is taken* w.r.t. N *and the observation law* π.

Related Studies Convergence analysis for the EM method in batch mode has been the focus of the classical work by Dempster et al. [1977], Wu et al. [1983], in which asymptotic convergence has been established; also see the recent work by Wang et al. [2015b], Xu et al. [2016b]. Several work has studied the convergence of stochastic EM with *fixed data*, e.g., Mairal [2015a] studied the asymptotic convergence to a stationary point, Chen et al. [2018] studied the local linear convergence of a variance reduced method by assuming that the iterates are bounded. On the other hand, the online EM method considered here, where a fresh sample is drawn at each iteration, has only been considered by a few work. Particularly, Cappé and Moulines [2009] showed the asymptotic convergence of the online EM method to a stationary point; Balakrishnan et al. [2017] analyzed non-asymptotic convergence for a variant of online EM method which requires a-priori the initial radius $\|\boldsymbol{\theta}_0 - \boldsymbol{\theta}^\star\|$, where $\boldsymbol{\theta}^\star$ is the optimal parameter. To our best knowledge, the rate results in Corollary 1 is new.

4.3.2 Policy Gradient for Average Reward over Infinite Horizon

There has been a growing interest in policy-gradient methods for model-free planning in Markov decision process; see [Sutton and Barto, 2018] and the references therein. Consider a finite Markov Decision Process (MDP) $(\mathsf{S}, \mathsf{A}, \mathrm{R}, \mathrm{P})$, where S is a finite set of spaces (state-space), A is a finite set of action (action-space), $\mathrm{R} : \mathsf{S} \times \mathsf{A} \to [0, \mathrm{R}_{\max}]$ is a reward function and P is the transition model, *i.e.,* given an action $a \in \mathsf{A}$, $\mathrm{P}^a = \{\mathrm{P}^a_{s,s'}\}$ is a matrix, $\mathrm{P}^a_{s,s'}$ is the probability of transiting from the sth state to the s'th state upon taking action a. The agent's decision is characterized by a parametric family of policies $\{\boldsymbol{\Pi}_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$: $\Pi_{\boldsymbol{\theta}}(a; s)$ which is the probability of taking action a when the current state is s (a semi-column is used to distinguish the random variables from parameters of the distribution). The state-action sequence $\{(S_t, A_t)\}_{t\geq 1}$ forms an MC with the transition matrix:

$$Q_{\boldsymbol{\theta}}((s,a);(s',a')) := \Pi_{\boldsymbol{\theta}}(a'; s')\,\mathrm{P}^a_{s,s'}\,, \tag{4.3.19}$$

where the above corresponds to the (s,a)th row, (s',a')th column of the matrix $\boldsymbol{Q}_{\boldsymbol{\theta}}$, and it denotes the transition probability from $(s,a) \in \mathsf{S} \times \mathsf{A}$ to $(s',a') \in \mathsf{S} \times \mathsf{A}$.

We assume that for each $\boldsymbol{\theta} \in \Theta$, the policy $\Pi_{\boldsymbol{\theta}}$ is ergodic, *i.e.,* $\boldsymbol{Q}_{\boldsymbol{\theta}}$ has a unique stationary distribution υ. Under this assumption, the *average reward* (or undiscounted reward) is given by

$$J(\boldsymbol{\theta}) := \textstyle\sum_{s,a} \upsilon(s,a)\, \mathrm{R}(s,a) \; . \tag{4.3.20}$$

The goal of the agent is to find a policy that maximizes the average reward over the class $\{\boldsymbol{\Pi}_{\boldsymbol{\theta}}\}_{\boldsymbol{\theta}\in\Theta}$. It can be verified [Sutton and Barto, 2018] that the gradient is evaluated by the limit:

$$\nabla J(\boldsymbol{\theta}) = \lim_{T\to\infty} \mathbb{E}_{\boldsymbol{\theta}}\big[\, \mathrm{R}(S_T, A_T) \textstyle\sum_{i=0}^{T-1} \nabla \log \Pi_{\boldsymbol{\theta}}(A_{T-i}; S_{T-i})\big] \; . \tag{4.3.21}$$

To approximate (4.3.21) with a numerically stable estimator, [Baxter and Bartlett, 2001] proposed the following gradient estimator. Let $\lambda \in [0,1)$ be a discount factor and T be sufficiently large, one has

$$\widehat{\nabla}_T J(\boldsymbol{\theta}) := \mathrm{R}(S_T, A_T) \textstyle\sum_{i=0}^{T-1} \lambda^i \nabla \log \Pi_{\boldsymbol{\theta}}(A_{T-i}; S_{T-i}) \approx \nabla J(\boldsymbol{\theta}) \; , \tag{4.3.22}$$

where $(S_1, A_1, \ldots, S_T, A_T)$ is a realization of state-action sequence generated by the policy $\Pi_{\boldsymbol{\theta}}$. This gradient estimator is *biased* and its bias is of order $O(1-\lambda)$ as the discount factor $\lambda \uparrow 1$. The approximation above leads to the following policy gradient method [Baxter and Bartlett, 2001]:

$$G_{n+1} = \lambda G_n + \nabla \log \Pi_{\boldsymbol{\theta}_n}(A_{n+1}; S_{n+1}) \; , \tag{4.3.23a}$$

$$\boldsymbol{\theta}_{n+1} = \boldsymbol{\theta}_n + \gamma_{n+1} G_{n+1}\, \mathrm{R}(S_{n+1}, A_{n+1}) \; . \tag{4.3.23b}$$

We focus on a linear parameterization of the policy in the exponential family (or soft-max):

$$\Pi_{\boldsymbol{\theta}}(a; s) = \big\{ \textstyle\sum_{a'\in\mathsf{A}} \exp\big(\langle \boldsymbol{\theta} \,|\, \boldsymbol{x}(s,a') - \boldsymbol{x}(s,a)\rangle \big) \big\}^{-1} \; , \tag{4.3.24}$$

where $\boldsymbol{x}(s,a) \in \mathbb{R}^d$ is a known feature vector. We make the following assumptions:

H4.10 *For all $s \in \mathsf{S}$, $a \in \mathsf{A}$, the feature vector $\boldsymbol{x}(s,a)$ and reward $\mathrm{R}(s,a)$ are bounded with $\|\boldsymbol{x}(s,a)\| \le \bar{b}, |\,\mathrm{R}(s,a)| \le \mathrm{R}_{\max}$.*

H4.11 *For all $\boldsymbol{\theta} \in \Theta$, the MC $\{(S_t, A_t)\}_{t\ge1}$, as governed by the transition matrix $\boldsymbol{Q}_{\boldsymbol{\theta}}$ [cf. (4.3.19)], is uniformly geometrically ergodic: there exists $\rho \in [0,1)$, $K_R < \infty$ such that, for all $n \ge 0$,*

$$\|\boldsymbol{Q}_{\boldsymbol{\theta}}^n - \mathbf{1}\boldsymbol{\upsilon}_{\boldsymbol{\theta}}^\top\| \le \rho^n K_R \; , \tag{4.3.25}$$

where $\boldsymbol{\upsilon}_{\boldsymbol{\theta}} \in \mathbb{R}_+^{|\mathsf{S}||\mathsf{A}|}$ is the stationary distribution of $\{(S_t, A_t)\}_{t\ge1}$. Moreover, there exists

$L_Q, L_\upsilon < \infty$ *such that for any* $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$,

$$\|\boldsymbol{\upsilon}_{\boldsymbol{\theta}} - \boldsymbol{\upsilon}_{\boldsymbol{\theta}'}\| \leq L_Q \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,, \quad \|\, \mathrm{J}^{\boldsymbol{\theta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}) - \mathrm{J}^{\boldsymbol{\theta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\theta}}}(\boldsymbol{\theta}')\| \leq L_\upsilon \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,, \tag{4.3.26}$$

where $\mathrm{J}^{\boldsymbol{\theta}}_{\boldsymbol{\upsilon}_{\boldsymbol{\theta}}}(\boldsymbol{\theta})$ *denotes the Jacobian of* $\boldsymbol{\upsilon}_{\boldsymbol{\theta}}$ w.r.t. $\boldsymbol{\theta}$.

Both H4.10 and H4.11 are regularity conditions on the MDP model that essentially hold as we focus on the finite state/action spaces setting. Under the uniform ergodicity assumption (4.3.25), the Lipschitz continuity conditions (4.3.26) can be implied using [Fort et al., 2011, Tadić and Doucet, 2017].

Our task is to verify that the policy gradient method (4.3.23) is an SA scheme with state-dependent Markovian noise [cf. Case 2 in Section 4.2]. To this end, we denote the joint state of this SA scheme as $X_n = (S_n, A_n, G_n) \in \mathsf{X} := \mathsf{S} \times \mathsf{A} \times \mathbb{R}^d$, and notice that $\{X_n\}_{n \geq 1}$ is a Markov chain. Adopting the same notation as in Section 4.2, the drift term and its mean field can be written as

$$H_{\boldsymbol{\theta}_n}(X_{n+1}) = G_{n+1}\, \mathrm{R}(S_{n+1}, A_{n+1}) \quad \text{with} \quad h(\boldsymbol{\theta}) = \lim_{T \to \infty} \mathbb{E}_{\tau_T \sim \Pi_{\boldsymbol{\theta}},\, S_1 \sim \overline{\Pi}_{\boldsymbol{\theta}}} \big[\widehat{\nabla}_T J(\boldsymbol{\theta}) \big] \,, \tag{4.3.27}$$

where $\widehat{\nabla}_T J(\boldsymbol{\theta})$ is defined in (4.3.22). Moreover, we let $P_{\boldsymbol{\eta}} : \mathsf{X} \times \mathcal{X} \to \mathbb{R}_+$ to be the Markov kernel associated with the MC $\{X_n\}_{n \geq 1}$. Observe that

Proposition 9 *Under H4.10, it holds for any* $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, $(s, a) \in \mathsf{S} \times \mathsf{A}$,

$$\|\nabla \log \Pi_{\boldsymbol{\theta}}(a; s)\| \leq 2\bar{b}, \quad \|\nabla \log \Pi_{\boldsymbol{\theta}}(a; s) - \nabla \log \Pi_{\boldsymbol{\theta}'}(a; s)\| \leq 8\bar{b}^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,. \tag{4.3.28}$$

Proof The proof is postponed to Appendix 4.7

Using the recursive update of (4.3.23a), we show that

$$\|G_n\| = \|\lambda G_{n-1} + \nabla \log \Pi_{\boldsymbol{\theta}}(A_n; S_n)\| \leq \lambda \|G_{n-1}\| + 2\bar{b} = \mathcal{O}(2\bar{b}\|G_0\|/(1-\lambda)) \,, \tag{4.3.29}$$

for any $n \geq 1$, which then implies that the stochastic update $H_{\boldsymbol{\theta}_n}(X_{n+1})$ in (4.3.23) is bounded since the reward is bounded using H4.10. The above proposition also implies that $h(\boldsymbol{\theta})$ is bounded for all $\boldsymbol{\theta} \in \Theta$. Therefore, the assumption H4.7 is satisfied.

Next, with a slight abuse of notation, we shall consider the compact state space $\mathsf{X} = \mathsf{S} \times \mathsf{A} \times \mathsf{G}$, with $\mathsf{G} = \{g \in \mathbb{R}^d : \|g\| \leq C_0 \bar{b}/(1-\lambda)\}$ and $C_0 \in [1, \infty)$, and analyze the policy gradient algorithm accordingly where $\{X_{n+1}\}_{n \geq 0}$ is in X.

Consider the following propositions whose proofs are adapted from [Fort et al., 2011, Tadić and Doucet, 2017]:

Proposition 10 *Under H4.10, H4.11, the following function is well-defined:*

$$\hat{H}_{\boldsymbol{\eta}}(x) = \sum_{t=0}^{\infty} \left\{ P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}) \right\} , \tag{4.3.30}$$

and satisfies Eq. (4.2.4). For all $x \in \mathsf{X}$, $(\boldsymbol{\eta}, \boldsymbol{\eta}') \in \Theta^2$, there exists constants $L_{PH}^{(0)}$, $L_{PH}^{(1)}$ where

$$\max\{\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x)\|, \|\hat{H}_{\boldsymbol{\eta}}(x)\|\} \leq L_{PH}^{(0)}, \quad \left\|P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'} \hat{H}_{\boldsymbol{\eta}'}(x)\right\| \leq L_{PH}^{(1)} \|\boldsymbol{\eta} - \boldsymbol{\eta}'\| . \tag{4.3.31}$$

Moreover, the constants are in the order of $L_{PH}^{(0)} = \mathcal{O}(\frac{1}{1-\max\{\rho,\lambda\}})$, $L_{PH}^{(1)} = \mathcal{O}(\frac{1}{1-\max\{\rho,\lambda\}})$.

Proposition 11 *Under H4.10, H4.11, the gradient $\nabla J(\boldsymbol{\theta})$ is Υ-Lipschitz continuous, where we defined $\Upsilon := \mathrm{R}_{\max} |\mathcal{S}||\mathcal{A}|$. Moreover, for any $\boldsymbol{\theta} \in \Theta$ and let $\Gamma := 2\bar{b}\, \mathrm{R}_{\max} K_R \frac{1}{(1-\rho)^2}$, it holds that*

$$(1-\lambda)^2\Gamma^2 + 2\left\langle \nabla J(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \right\rangle \geq \|h(\boldsymbol{\theta})\|^2, \ \|\nabla J(\boldsymbol{\theta})\| \leq \|h(\boldsymbol{\theta})\| + (1-\lambda)\Gamma . \tag{4.3.32}$$

Proof The proofs are postponed to Appendix 4.7

Proposition 10 verifies H4.5 and H4.6 for the policy gradient algorithm, while Proposition 11 implies H4.1 [with $c_0 = (1-\lambda)^2\Gamma^2$, $c_1 = 2$], H4.2 [with $d_0 = (1-\lambda)\Gamma$, $d_1 = 1$], H4.3 [with $L = \Upsilon$]. As such, applying Theorem 4 shows that

Corollary 2 *Under H4.10, H4.11 and set $\gamma_k = (2c_1 L(1+C_h)\sqrt{k})^{-1}$. For any $n \in \mathbb{N}$, let $N \in \{0, ..., n\}$ be an independent discrete r.v. distributed according to (4.2.1), the policy gradient algorithm (4.3.23) finds a policy such that*

$$\mathbb{E}[\|\nabla J(\boldsymbol{\theta}_N)\|^2] = \mathcal{O}\Big((1-\lambda)^2\Gamma^2 + \log n/\sqrt{n}\Big), \tag{4.3.33}$$

where $J(\cdot)$ is defined in (4.3.20) and the expectation is taken w.r.t. *N and action-state pairs (A_n, S_n).*

Related Studies The convergence of policy gradient method is typically studied for the *episodic* setting where the goal is to maximize the total reward over a *finite horizon*. The REINFORCE algorithm [Williams, 1992] has been analyzed as an SG method with *unbiased* gradient estimate in [Sutton et al., 2000], which proved an asymptotic convergence condition. A recent work [Papini et al., 2018] combined the variance reduction technique with the REINFORCE algorithm.

The *infinite horizon* setting is more challenging. To our best knowledge, the first asymptotically convergent policy gradient method is the actor-critic algorithm by Konda and Tsitsiklis [2003] which is extended to off-policy learning in [Degris et al., 2012]. The analysis are based on the theory of two time-scales SA, which relies on controlling the ratio between the two set of step sizes used [Borkar, 1997]. On the other hand, the algorithm which we have studied was a direct policy gradient method proposed by Baxter and Bartlett [2001], whose asymptotic convergence was proven only recently by Tadić and Doucet [2017]. In comparison, our Corollary 2 provides the first non-asymptotic convergence for the policy gradient method. Of related interest, it is worthwhile to mention that [Abbasi-Yadkori et al., 2018, Fazel et al., 2018] have studied the global convergence for average reward maximization under the linear quadratic regulator setting where the state transition can be characterized by a linear dynamics and the reward is a quadratic function.

4.4 Conclusion

In this paper, we analyze under mild assumptions a general SA scheme with either *zero-mean* [cf. Case 1] or *state-dependent/controlled Markovian* [cf. Case 2] noise. We establish a novel *non-asymptotic* convergence analysis of this procedure without assuming convexity of the Lyapunov function. In both cases, our results highlight a convergence rate of order $\mathcal{O}(\log(n)/\sqrt{n})$ under conservative assumptions. We verify our findings on two applications of growing interest: the online EM for learning an exponential family distribution (e.g., Gaussian Mixture Model) and the policy gradient method for maximizing an average reward.

Appendices to Online Optimization of Non-convex Problems

4.5 Proofs of Section 4.2.1

4.5.1 Proof of Lemma 3

Lemma *Assume H4.1, H4.3. Then, for all $n \geq 1$, it holds that:*

$$\begin{aligned}&\textstyle\sum_{k=0}^{n} \frac{\gamma_{k+1}}{c_1}(1 - c_1 L \gamma_{k+1}) h_k \\ &\leq V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}_{n+1}) + L \textstyle\sum_{k=0}^{n} \gamma_{k+1}^2 \|\boldsymbol{e}_{k+1}\|^2 + \sum_{k=0}^{n} \gamma_{k+1}(c_1^{-1} c_0 - \langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1} \rangle) \,.\end{aligned} \tag{4.5.1}$$

Proof As the Lyapunov function $V(\boldsymbol{\theta})$ is L smooth [cf. H4.3], we obtain:

$$\begin{aligned}V(\boldsymbol{\theta}_{k+1}) &\leq V(\boldsymbol{\theta}_k) - \gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, H_{\boldsymbol{\theta}_k}(X_{k+1}) \rangle + \frac{L \gamma_{k+1}^2}{2} \|H_{\boldsymbol{\theta}_k}(X_{k+1})\|^2 \\ &\leq V(\boldsymbol{\theta}_k) - \gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, h(\boldsymbol{\theta}_k) + \boldsymbol{e}_{k+1} \rangle + L \gamma_{k+1}^2 (\|h(\boldsymbol{\theta}_k)\|^2 + \|\boldsymbol{e}_{k+1}\|^2) \,.\end{aligned} \tag{4.5.2}$$

The above implies that

$$\begin{aligned}\gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, h(\boldsymbol{\theta}_k) \rangle &\leq V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k+1}) - \gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1} \rangle \\ &\quad + L \gamma_{k+1}^2 (\|h(\boldsymbol{\theta}_k)\|^2 + \|\boldsymbol{e}_{k+1}\|^2) \,.\end{aligned} \tag{4.5.3}$$

Using H4.1, $\langle \nabla V(\boldsymbol{\theta}_k) \,|\, h(\boldsymbol{\theta}_k) \rangle \geq \frac{1}{c_1}(h_k - c_0)$ and rearranging terms, we obtain

$$\begin{aligned}\frac{\gamma_{k+1}}{c_1}(1 - c_1 L \gamma_{k+1}) h_k &\leq V(\boldsymbol{\theta}_k) - V(\boldsymbol{\theta}_{k+1}) - \gamma_{k+1} \langle \nabla V(\boldsymbol{\theta}_k) \,|\, \boldsymbol{e}_{k+1} \rangle \\ &\quad + L \gamma_{k+1}^2 \|\boldsymbol{e}_{k+1}\|^2 + \frac{c_0}{c_1} \gamma_{k+1} \,.\end{aligned} \tag{4.5.4}$$

Summing up both sides from $k = 0$ to $k = n$ gives the conclusion (4.2.12). ■

4.5.2 Proof of Lemma 4

Lemma *Assume H4.1–H4.3, H4.5–H4.7 and the step sizes satisfy (4.2.6). Then:*

$$\mathbb{E}\left[-\sum_{k=0}^{n}\gamma_{k+1}\left\langle\nabla V(\boldsymbol{\theta}_k)\,|\,\boldsymbol{e}_{k+1}\right\rangle\right] \le C_h\sum_{k=0}^{n}\gamma_{k+1}^2\mathbb{E}[\|h(\boldsymbol{\theta}_k)\|^2] + C_\gamma\sum_{k=0}^{n}\gamma_{k+1}^2 + C_{0,n}\ , \quad (4.5.5)$$

where C_h, C_γ and $C_{0,n}$ are defined in (4.2.8), (4.2.9), (4.2.10).

Proof Under H4.5, H4.7, for any $\boldsymbol{\theta} \in \Theta$ there exists a bounded, measurable function $x \to \hat{H}_{\boldsymbol{\theta}}(x)$ such that the Poisson equation holds:

$$\boldsymbol{e}_{n+1} = H_{\boldsymbol{\theta}_n}(X_{n+1}) - h(\boldsymbol{\theta}_n) = \hat{H}_{\boldsymbol{\theta}_n}(X_{n+1}) - P_{\boldsymbol{\eta}_n}\hat{H}_{\boldsymbol{\theta}_n}(X_{n+1})\ . \quad (4.5.6)$$

The inner product on the left hand side of (4.2.17) can thus be decomposed as

$$\mathbb{E}\left[-\sum_{k=0}^{n}\gamma_{k+1}\left\langle\nabla V(\boldsymbol{\theta}_k)\,|\,\boldsymbol{e}_{k+1}\right\rangle\right] = \mathbb{E}[A_1 + A_2 + A_3 + A_4 + A_5]\ , \quad (4.5.7)$$

with

$$\begin{aligned}
A_1 &:= -\sum_{k=1}^{n}\gamma_{k+1}\left\langle\nabla V(\boldsymbol{\theta}_k)\,|\,\hat{H}_{\boldsymbol{\theta}_k}(X_{k+1}) - P_{\boldsymbol{\eta}_k}\hat{H}_{\boldsymbol{\theta}_k}(X_k)\right\rangle\ ,\\
A_2 &:= -\sum_{k=1}^{n}\gamma_{k+1}\left\langle\nabla V(\boldsymbol{\theta}_k)\,|\,P_{\boldsymbol{\eta}_k}\hat{H}_{\boldsymbol{\theta}_k}(X_k) - P_{\boldsymbol{\eta}_{k-1}}\hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k)\right\rangle\ ,\\
A_3 &:= -\sum_{k=1}^{n}\gamma_{k+1}\left\langle\nabla V(\boldsymbol{\theta}_k) - \nabla V(\boldsymbol{\theta}_{k-1})\,|\,P_{\boldsymbol{\eta}_{k-1}}\hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k)\right\rangle\ ,\\
A_4 &:= -\sum_{k=1}^{n}(\gamma_{k+1} - \gamma_k)\left\langle\nabla V(\boldsymbol{\theta}_{k-1})\,|\,P_{\boldsymbol{\eta}_{k-1}}\hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k)\right\rangle\ ,\\
A_5 &:= -\gamma_1\left\langle\nabla V(\boldsymbol{\theta}_0)\,|\,\hat{H}_{\boldsymbol{\theta}_0}(X_1)\right\rangle + \gamma_{n+1}\left\langle\nabla V(\boldsymbol{\theta}_n)\,|\,P_{\boldsymbol{\eta}_n}\hat{H}_{\boldsymbol{\theta}_n}(X_{n+1})\right\rangle\ .
\end{aligned}$$

For A_1, we note that $\hat{H}_{\boldsymbol{\theta}_k}(X_{k+1}) - P_{\boldsymbol{\eta}_k}\hat{H}_{\boldsymbol{\theta}_k}(X_k)$ is a martingale difference sequence [cf. (4.1.2)] and therefore we have $\mathbb{E}[A_1] = 0$ by taking the total expectation.

For A_2, applying the Cauchy-Schwarz inequality and (4.2.5), we have

$$\begin{aligned}
A_2 &\leq L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1} \|\nabla V(\boldsymbol{\theta}_k)\| \|\boldsymbol{\theta}_k - \boldsymbol{\theta}_{k-1}\| \\
&= L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \|\nabla V(\boldsymbol{\theta}_k)\| \|H_{\boldsymbol{\theta}_{k-1}}(X_k)\| \\
&\overset{(a)}{\leq} L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k (d_0 + d_1 \|h(\boldsymbol{\theta}_k)\|)(\|h(\boldsymbol{\theta}_{k-1})\| + \sigma) \\
&\overset{(b)}{\leq} L_{PH}^{(1)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \Big(d_0\sigma + d_0 \|h(\boldsymbol{\theta}_{k-1})\| + d_1\sigma \|h(\boldsymbol{\theta}_k)\| + d_1 \|h(\boldsymbol{\theta}_k)\| \|h(\boldsymbol{\theta}_{k-1})\| \Big) ,
\end{aligned} \tag{4.5.8}$$

where (a) is due to H4.2 on the norm of $\nabla V(\boldsymbol{\theta}_k)$ and H4.7 on the norm of $\boldsymbol{e}_k$, (b) is obtained by expanding the scalar product. Using the inequality $\|h(\boldsymbol{\theta}_n)\| \leq 1 + \|h(\boldsymbol{\theta}_n)\|^2$ and $2\|h(\boldsymbol{\theta}_k)\|\|h(\boldsymbol{\theta}_{k-1})\| \leq \|h(\boldsymbol{\theta}_k)\|^2 + \|h(\boldsymbol{\theta}_{k-1}\|^2$, we obtain:

$$A_2 \leq L_{PH}^{(1)} \left((d_0 + d_0\sigma + d_1\sigma) \sum_{k=1}^{n} \gamma_k^2 + (d_0 + \frac{d_1}{2} + a d_1 \sigma + \frac{a d_1}{2}) \sum_{k=0}^{n} \gamma_{k+1}^2 \|h(\boldsymbol{\theta}_k)\|^2 \right). \tag{4.5.9}$$

For A_3, we obtain

$$\begin{aligned}
A_3 &\overset{(a)}{\leq} L \sum_{k=1}^{n} \gamma_{k+1}\gamma_k \|H_{\boldsymbol{\theta}_{k-1}}(X_k)\| \|P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k)\| \\
&\overset{(b)}{\leq} L L_{PH}^{(0)} \sum_{k=1}^{n} \gamma_{k+1}\gamma_k (\|h(\boldsymbol{\theta}_{k-1})\| + \sigma) \\
&\leq L L_{PH}^{(0)} \left((1+\sigma) \sum_{k=1}^{n} \gamma_k^2 + \sum_{k=1}^{n} \gamma_k^2 \|h(\boldsymbol{\theta}_{k-1})\|^2 \right) ,
\end{aligned} \tag{4.5.10}$$

where (a) uses H4.3, (b) uses $H_{\boldsymbol{\theta}_{k-1}}(X_k) = h(\boldsymbol{\theta}_{k-1}) + \boldsymbol{e}_k$ and H4.6.

For A_4, we have

$$\begin{aligned}
A_4 &\leq \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| (d_0 + d_1 \|h(\boldsymbol{\theta}_{k-1})\|) \|P_{\boldsymbol{\eta}_{k-1}} \hat{H}_{\boldsymbol{\theta}_{k-1}}(X_k)\| \\
&\overset{(a)}{\leq} L_{PH}^{(0)} \left((d_0 + 1) \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| + d_1 \sum_{k=1}^{n} |\gamma_{k+1} - \gamma_k| \|h(\boldsymbol{\theta}_{k-1})\|^2 \right) \\
&\overset{(b)}{=} L_{PH}^{(0)} \left((d_0 + 1)(\gamma_1 - \gamma_{n+1}) + a' d_1 \sum_{k=1}^{n} \gamma_k^2 \|h(\boldsymbol{\theta}_{k-1})\|^2 \right) ,
\end{aligned} \tag{4.5.11}$$

where (a) is again an application of H4.6, and (b) uses the assumptions on step size

$\gamma_{k+1} \leq \gamma_k$, $\gamma_k - \gamma_{k+1} \leq a'\gamma_k^2$. Finally, for A_5, we obtain

$$\begin{aligned} A_5 &\overset{(a)}{\leq} \gamma_1(d_0 + d_1\|h(\boldsymbol{\theta}_0)\|)L_{PH}^{(0)} + \gamma_{n+1}(d_0 + d_1\|h(\boldsymbol{\theta}_n)\|)L_{PH}^{(0)} \\ &\overset{(b)}{\leq} L_{PH}^{(0)}\Big(d_0\{\gamma_1 + \gamma_{n+1}\} + 2d_1 + d_1\{\gamma_1^2\|h(\eta_0)\|^2 + \gamma_{n+1}^2\|h(\eta_n)\|^2\}\Big) \\ &\leq L_{PH}^{(0)}\Big(d_0\{\gamma_1 + \gamma_{n+1}\} + 2d_1 + d_1\sum_{k=0}^{n}\gamma_{k+1}^2\|h(\boldsymbol{\theta}_k)\|^2\Big)\,, \end{aligned} \tag{4.5.12}$$

where (a) is an application of H4.2 and H4.6, and (b) uses $a \leq 1 + a^2$. Gathering the relevant terms and taking expectations conclude the proof of this lemma. ■

4.5.3 Proof of Lemma 2

Lemma *Consider the SA scheme (4.1.1) with $h(\boldsymbol{\theta}) = \nabla V(\boldsymbol{\theta})$. There exists a Lyapunov function $V(\boldsymbol{\theta})$ satisfying H4.3 and a noise sequence $\{\boldsymbol{e}_n\}_{n\geq 1}$ satisfying H4.4-H4.7 such that for any $n \geq 1$,*

$$\mathbb{E}[\|h(\boldsymbol{\theta}_N)\|^2] \geq \frac{\mathbb{E}[V(\boldsymbol{\theta}_0) - V(\boldsymbol{\theta}_{n+1})] + C_{\mathsf{lb}}\sum_{k=0}^{n}\gamma_{k+1}^2}{\sum_{k=0}^{n}\gamma_{k+1}} \tag{4.5.13}$$

where N is distributed according to (4.2.1), and $C_{\mathsf{lb}} > 0$ is some constant independent of n.

Proof Our proof is achieved through constructing the Lyapunov and mean field function below. Consider a scalar parameter $\eta \in \mathbb{R}$ and set $V(\eta)$ to be a μ-strongly convex and L-smooth function, where $0 < \mu \leq L < \infty$. Also, the mean field is set as

$$h(\eta) = V'(\eta)\,. \tag{4.5.14}$$

Consider the following SA scheme (4.1.1) defined on the mean field h as:

$$\eta_{k+1} = \eta_k - \gamma_{k+1}\big(h(\eta_k) + e_{k+1}\big)\,, \tag{4.5.15}$$

where e_k is i.i.d. and uniformly distributed on $[-\varepsilon, \varepsilon]$.

Clearly, the SA scheme (4.5.15) satisfies H4.1-H4.3 as we have set $V'(\eta) = h(\eta)$. The noise sequence is i.i.d. satisfying H4.4-H4.7. As V is μ-strongly convex, it can be shown

$$V(\eta_{k+1}) \geq V(\eta_k) - \gamma_{k+1}V'(\eta_k)\big(h(\eta_k) + e_{k+1}\big) + \gamma_{k+1}^2\frac{\mu}{2}\big(h(\eta_k) + e_{k+1}\big)^2\,. \tag{4.5.16}$$

Now by construction, we have $\mathbb{E}[e_{k+1}V'(\eta_k)|\mathcal{F}_k] = 0$, $\mathbb{E}[\big(h(\eta_k) + e_{k+1}\big)^2|\mathcal{F}_k] \geq \frac{1}{3}\varepsilon^2$. Taking

the total expectation on both sides gives

$$\mathbb{E}[V(\eta_{k+1})] \geq \mathbb{E}[V(\eta_k)] - \gamma_{k+1} h^2(\eta_k) + \gamma_{k+1}^2 \frac{\mu \varepsilon^2}{6} \ . \tag{4.5.17}$$

Denote $C_{\mathsf{lb}} := \frac{\mu \varepsilon^2}{6}$. Using (4.2.1), we observe

$$\mathbb{E}[|h(\eta_N)|^2] = \frac{1}{\sum_{k=0}^n \gamma_{k+1}} \sum_{k=0}^n \gamma_{k+1} \mathbb{E}[|h(\eta_k)|^2] \geq \frac{\mathbb{E}[V(\eta_0) - V(\eta_{n+1})] + C_{\mathsf{lb}} \sum_{k=0}^n \gamma_{k+1}^2}{\sum_{k=0}^n \gamma_{k+1}}. \tag{4.5.18}$$

This completes the proof of the lower bound.

4.6 Proofs for the ro-EM Method

4.6.1 Proof of Proposition 5

Proposition *Assume H5.4. Then*

- *If* $h(\boldsymbol{s}^\star) = \mathbf{0}$ *for some* $\boldsymbol{s}^\star \in \mathsf{S}$, *then* $\nabla_{\boldsymbol{\theta}} \mathrm{KL}\,(\pi, g_{\boldsymbol{\theta}^\star}) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\boldsymbol{\theta}^\star) = \mathbf{0}$ *with* $\boldsymbol{\theta}^\star = \overline{\boldsymbol{\theta}}(\boldsymbol{s}^\star)$.
- *If* $\nabla_{\boldsymbol{\theta}} \mathrm{KL}\,(\pi, g_{\boldsymbol{\theta}^\star}) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\boldsymbol{\theta}^\star) = \mathbf{0}$ *for some* $\boldsymbol{\theta}^\star \in \Theta$ *then* $\boldsymbol{s}^\star = \mathbb{E}_\pi[S(Y, \boldsymbol{\theta}^\star)]$.

Proof We have

$$\nabla_{\boldsymbol{\theta}} \mathrm{KL}\,(\pi, g(\cdot; \boldsymbol{\theta})) = -\nabla_{\boldsymbol{\theta}} \mathbb{E}_\pi\big[\log g(Y; \boldsymbol{\theta})\big] = -\mathbb{E}_\pi\big[\nabla_{\boldsymbol{\theta}} \log g(Y; \boldsymbol{\theta})\big] \ , \tag{4.6.1}$$

where the last equality assumes that we can exchange integration with differentiation. Furthermore, using the Fisher's identity [Douc et al., 2014], it holds for any $y \in \mathsf{Y}$ that

$$\nabla_{\boldsymbol{\theta}} \log g(y; \boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) + J_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})\, \overline{\boldsymbol{s}}(y; \boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) + J_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})\, \mathbb{E}_{\boldsymbol{\theta}}\big[S(\boldsymbol{X}) | Y = y\big] \ . \tag{4.6.2}$$

Therefore, for any $\boldsymbol{s}$, it holds that

$$\begin{aligned}
&\nabla_{\boldsymbol{\theta}} \mathrm{KL}\left(\pi, g(\cdot; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\right) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) = \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) - J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\, \mathbb{E}_\pi\big[\overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big] \\
&\overset{(a)}{=} J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\boldsymbol{s} - \mathbb{E}_\pi\big[\overline{\boldsymbol{s}}(Y; \overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big]\Big) \overset{(b)}{=} J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\, h(\boldsymbol{s}) \ .
\end{aligned} \tag{4.6.3}$$

where we have used the assumption H5.4 in (a) and the definition of $h(\boldsymbol{s})$ in (b). The conclusion follows directly from the identity (4.6.3) since $J_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$ is full rank. ■

4.6.2 Proof of Proposition 6

Proposition *Assume H5.4. Then, for* $\boldsymbol{s} \in \mathsf{S}$,

$$\nabla_{\boldsymbol{s}} V(\boldsymbol{s}) = \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})\Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top} h(\boldsymbol{s}) \,. \tag{4.6.4}$$

Proof Using chain rule and H5.4, we obtain

$$\begin{aligned} \nabla_{\boldsymbol{s}} V(\boldsymbol{s}) &= \mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s})^{\top}\Big(\nabla_{\boldsymbol{\theta}}\,\mathrm{KL}\left(\pi, g(\cdot;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\right) + \nabla_{\boldsymbol{\theta}}\,\mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big) \\ &= \mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s})^{\top}\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top} h(\boldsymbol{s}) \,, \end{aligned} \tag{4.6.5}$$

where the last equality uses the identity in (4.6.3). Consider the following vector map:

$$\boldsymbol{s} \to \nabla_{\boldsymbol{\theta}}\psi(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \nabla_{\boldsymbol{\theta}}\,\mathrm{R}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^{\top}\,\boldsymbol{s} \,. \tag{4.6.6}$$

Taking the gradient of the above map *w.r.t.* $\boldsymbol{s}$ and note that the map is constant for all $\boldsymbol{s} \in \mathsf{S}$, we show that:

$$\mathbf{0} = -\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s})) + \Big(\underbrace{\nabla_{\boldsymbol{\theta}}^2(\psi(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) \,|\, \boldsymbol{s}\rangle\,)}_{=\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})}\big|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\boldsymbol{s})}\Big)\,\mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s}) \,. \tag{4.6.7}$$

This implies $\mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\boldsymbol{s}}(\boldsymbol{s}) = \big(\mathrm{H}_{\ell}^{\boldsymbol{\theta}}(\boldsymbol{s};\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1}\,\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$. Substituting into (5.9.1) yields the conclusion. ■

4.6.3 Proof of Proposition 7

Proposition *Under H4.9, it holds that* $\mathbb{E}[\|\overline{s}(Y_{n+1};\hat{\boldsymbol{\theta}}_n) - \hat{\boldsymbol{s}}_n\|^2|\mathcal{F}_n] \leq 2M\overline{Y}^2$ *for all* $n \geq 0$.

Proof From (4.3.7), we note that the error term is given by

$$\boldsymbol{e}_{n+1} = H_{\hat{\boldsymbol{s}}_n}(Y_{n+1}) - h(\hat{\boldsymbol{s}}_n) = \begin{pmatrix} \mathbb{E}_{Y_{n+1}\sim\pi}[\overline{\boldsymbol{s}}_n^{(1)}|\mathcal{F}_n] - \overline{\boldsymbol{s}}_n^{(1)} \\ \mathbb{E}_{Y_{n+1}\sim\pi}[\overline{\boldsymbol{s}}_n^{(2)}|\mathcal{F}_n] - \overline{\boldsymbol{s}}_n^{(2)} \\ \mathbb{E}_{Y_{n+1}\sim\pi}[\overline{s}_n^{(3)}|\mathcal{F}_n] - \overline{s}_n^{(3)} \end{pmatrix}. \tag{4.6.8}$$

Obviously, it holds that $\mathbb{E}[\boldsymbol{e}_{n+1}|\mathcal{F}_n] = \mathbf{0}$. Furthermore, for all $m \in \{1,\ldots,M-1\}$, the mth element of the first block in $\boldsymbol{e}_{n+1}$ has a bounded conditional variance

$$\mathbb{E}\Big[\big|\mathbb{E}_{Y_{n+1}\sim\pi}[\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)] - \omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] \leq 1 \,. \tag{4.6.9}$$

For the second block in $\boldsymbol{e}_{n+1}$, the conditional variance of its mth element is

$$\begin{aligned}
&\mathbb{E}\Big[\big|\mathbb{E}_{Y_{n+1}\sim\pi}[Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)] - Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] \\
&= \mathbb{E}\Big[\big|Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] - \big|\mathbb{E}_{Y_{n+1}\sim\pi}[Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)]\big|^2 \\
&\le \mathbb{E}\Big[\big|Y_{n+1}\omega_m(Y_{n+1};\hat{\boldsymbol{\theta}}_n)\big|^2\Big] \le \mathbb{E}\big[(Y_{n+1})^2\big] \le \overline{Y}^2.
\end{aligned} \tag{4.6.10}$$

Lastly, we also have $\mathbb{E}[|\mathbb{E}_{Y_{n+1}\sim\pi}[\bar{s}_n^{(3)}|\mathcal{F}_n] - \bar{s}_n^{(3)}|^2] \le \overline{Y}^2$. Therefore, we conclude that $\mathbb{E}[\|\boldsymbol{e}_{n+1}\|^2|\mathcal{F}_n] \le M - 1 + M\overline{Y}^2 < \infty$. ∎

4.6.4 Proof of Proposition 8

Proposition *Under H4.9 and the regularizer (7.3.6) set with $\epsilon > 0$, then for all $(\boldsymbol{s}, \boldsymbol{s}') \in \mathsf{S}^2$, there exists positive constants υ, Υ, Ψ such that:*

$$\langle \nabla V(\boldsymbol{s}) \,|\, h(\boldsymbol{s})\rangle \ge \upsilon \,\|h(\boldsymbol{s})\|^2, \quad \|\nabla V(\boldsymbol{s})\| \le \Upsilon\|h(\boldsymbol{s})\|, \quad \|\nabla V(\boldsymbol{s}) - \nabla V(\boldsymbol{s}')\| \le \Psi\|\boldsymbol{s} - \boldsymbol{s}'\|. \tag{4.6.11}$$

Proof We first check that H5.4 is satisfied under H4.9. In particular, one observes that when $\boldsymbol{s} \in \mathsf{S} = \Delta_{M-1}\times[-\overline{Y}, \overline{Y}]^M$, the M-step update (7.3.7) is the unique solution satisfying the stationary condition of the minimization problem (5.2.5) and $\overline{\boldsymbol{\theta}}(\boldsymbol{s}) \in \mathcal{C}$.

As H5.4 is satisfied, applying Proposition 6 shows that the gradient of the Lyapunov function is

$$\nabla V(\boldsymbol{s}) = \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})\}\Big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top h(\boldsymbol{s})\,. \tag{4.6.12}$$

Using (7.3.3), we observe that for any given $\boldsymbol{\theta} \in \mathcal{C}$, the Jacobian of ϕ and the Hessian of $\ell(\boldsymbol{s}, \boldsymbol{\theta})$ are given by

$$\mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{1-\sum_{m=1}^{M-1}\omega_m}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{\mathbf{1}}{\boldsymbol{\omega}}) & -\mathrm{Diag}(\boldsymbol{\mu}) & \mu_M\mathbf{1} \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{pmatrix}, \tag{4.6.13}$$

$$\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s}, \boldsymbol{\theta}) = \begin{pmatrix} \frac{1+\epsilon-\sum_{m=1}^{M-1}s_m^{(1)}}{(1-\sum_{m=1}^{M-1}\omega_m)^2}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}{\boldsymbol{\omega}^2}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\boldsymbol{s}^{(1)} + \epsilon\mathbf{1}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 + \epsilon - \sum_{m=1}^{M-1} s_m^{(1)} \end{pmatrix},$$

where we have denoted $\frac{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}{\boldsymbol{\omega}^2}$ as the $(M-1)$-vector $(\frac{s_1^{(1)}+\epsilon}{\omega_1^2}, \ldots, \frac{s_{M-1}^{(1)}+\epsilon}{\omega_{M-1}^2})$. Let us define $\boldsymbol{J}_{11}, \boldsymbol{H}_{11}$ as the top-left matrices in the above, evaluated at $\overline{\boldsymbol{\theta}}(\boldsymbol{s})$, as follows

$$\boldsymbol{J}_{11} := \frac{1}{1 - \frac{\mathbf{1}^\top(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})}{1+\epsilon M}}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{1+\epsilon M}{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}) \tag{4.6.14}$$

$$\boldsymbol{H}_{11} := \frac{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}{(1-\frac{\mathbf{1}^\top(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})}{1+\epsilon M})^2}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{(1+\epsilon M)^2}{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}). \tag{4.6.15}$$

When $\epsilon > 0$, the above matrices, $\boldsymbol{J}_{11}$ and $\boldsymbol{H}_{11}$, are full rank and bounded if $\boldsymbol{s} \in \mathsf{S}$.

The matrix product $\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s},\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top$ can hence be expressed as an outer product

$$\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s},\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top = \boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top\,, \tag{4.6.16}$$

with

$$\begin{aligned}\boldsymbol{\mathcal{J}}(\boldsymbol{s}) &:= \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\begin{pmatrix} \boldsymbol{H}_{11}^{-\frac{1}{2}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}} \end{pmatrix} \\ &= \begin{pmatrix} \boldsymbol{J}_{11}\boldsymbol{H}_{11}^{-\frac{1}{2}} & -\mathrm{Diag}(\frac{\boldsymbol{s}^{(2)}}{(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})^{\frac{3}{2}}}) & \frac{s^{(3)}-\mathbf{1}^\top\boldsymbol{s}^{(2)}}{(1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)})^{\frac{3}{2}}}\mathbf{1} \\ \mathbf{0} & \mathrm{Diag}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}} \end{pmatrix}.\end{aligned} \tag{4.6.17}$$

Under H4.9 and using the above structured form, it can be verified that $\boldsymbol{\mathcal{J}}(\boldsymbol{s})$ is a bounded and full rank matrix. As such, for all $\boldsymbol{s} \in \mathsf{S}$, there exists $\upsilon > 0$ such that

$$\langle \nabla V(\boldsymbol{s}) \,|\, h(\boldsymbol{s})\rangle = \Big\langle \boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top h(\boldsymbol{s}) \,|\, h(\boldsymbol{s})\Big\rangle \geq \upsilon\,\|h(\boldsymbol{s})\|^2\,. \tag{4.6.18}$$

The second part in (4.3.17) can be verified by observing that $\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\Big(\mathrm{H}_\ell^{\boldsymbol{\theta}}(\boldsymbol{s};\boldsymbol{\theta})\}\Big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))^\top$ is bounded due to H4.9.

For the third part in (4.3.17), again from (4.6.12) we obtain:

$$\nabla V(\boldsymbol{s}) = \boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top h(\boldsymbol{s})\,. \tag{4.6.19}$$

From (5.13.11), it can be seen that $\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top$ is Lipschitz continuous in $\boldsymbol{s}$ and bounded, *i.e.,* there exists constants $L_J, C_J < \infty$ such that

$$\|\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top - \boldsymbol{\mathcal{J}}(\boldsymbol{s}')\boldsymbol{\mathcal{J}}(\boldsymbol{s}')^\top\| \leq L_J\|\boldsymbol{s}-\boldsymbol{s}'\|,\quad \|\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top\| \leq C_J,\ \forall\ \boldsymbol{s},\boldsymbol{s}' \in \mathsf{S}\,. \tag{4.6.20}$$

For example, the above can be checked by observing that the Hessian (*w.r.t.* $\boldsymbol{s}$) of each entry in $\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top$ is bounded for $\boldsymbol{s} \in \mathsf{S}$. On the other hand, the mean field $h(\boldsymbol{s})$ satisfies,

$$\begin{aligned}\|h(\boldsymbol{s})-h(\boldsymbol{s}')\| &= \|\boldsymbol{s}-\boldsymbol{s}'+\mathbb{E}_{Y\sim\pi}\big[\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}'))-\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\big]\| \\ &\overset{(a)}{\leq} \|\boldsymbol{s}-\boldsymbol{s}'\| + \mathbb{E}_{Y\sim\pi}\big[\|\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}'))-\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\|\big]\,,\end{aligned} \tag{4.6.21}$$

where (a) uses the triangular inequality and the Jensen's inequality. Moreover, we observe

$$\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}'))-\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))=\begin{pmatrix}\widetilde{\boldsymbol{\omega}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s})')-\widetilde{\boldsymbol{\omega}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\\ Y(\widetilde{\boldsymbol{\omega}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s})')-\widetilde{\boldsymbol{\omega}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s})))\\ 0\end{pmatrix}, \tag{4.6.22}$$

where $\widetilde{\boldsymbol{\omega}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$ is a collection of the $M-1$ terms $\widetilde{\omega}_m(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))$, $m=1,\ldots,M-1$ [cf. (7.3.4)]. Observe that

$$\widetilde{\omega}_m(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))=\frac{\frac{s_m^{(1)}+\epsilon}{1+\epsilon M}\exp(-\frac{1}{2}(Y-\frac{s_m^{(2)}}{s_m^{(1)}+\epsilon})^2)}{\sum_{j=1}^{M}\frac{s_j^{(1)}+\epsilon}{1+\epsilon M}\exp(-\frac{1}{2}(Y-\frac{s_j^{(2)}}{s_j^{(1)}+\epsilon})^2)}. \tag{4.6.23}$$

Under H4.9 and the condition that $\boldsymbol{s}\in\mathsf{S}$, *i.e.,* a compact set, there exists $L_\omega<\infty$ such that

$$|\widetilde{\omega}_m(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))-\widetilde{\omega}_m(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}'))|^2\le L_\omega^2\|\boldsymbol{s}-\boldsymbol{s}'\|^2, \tag{4.6.24}$$

for all $m=1,\ldots,M-1$. Consequently, again using H4.9, we have

$$\|\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}'))-\overline{\boldsymbol{s}}(Y;\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\|\le(M-1)(1+\overline{Y})L_\omega\|\boldsymbol{s}-\boldsymbol{s}'\|, \tag{4.6.25}$$

and we have $\|h(\boldsymbol{s})-h(\boldsymbol{s}')\|\le L_h\|\boldsymbol{s}-\boldsymbol{s}'\|$ for some $L_h<\infty$. It can also be shown easily that $\|h(\boldsymbol{s})\|\le C_h$ for all $\boldsymbol{s}\in\mathsf{S}$. Finally, we observe the following chain:

$$\begin{aligned}&\|\nabla V(\boldsymbol{s})-\nabla V(\boldsymbol{s}')\|=\|\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top h(\boldsymbol{s})-\boldsymbol{\mathcal{J}}(\boldsymbol{s}')\boldsymbol{\mathcal{J}}(\boldsymbol{s}')^\top h(\boldsymbol{s}')\|\\&=\|\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top(h(\boldsymbol{s})-h(\boldsymbol{s}'))+(\boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top-\boldsymbol{\mathcal{J}}(\boldsymbol{s}')\boldsymbol{\mathcal{J}}(\boldsymbol{s}')^\top)h(\boldsymbol{s}')\|\\&\le(L_hC_J+L_JC_h)\|\boldsymbol{s}-\boldsymbol{s}'\|,\end{aligned} \tag{4.6.26}$$

which concludes our proof. ■

4.7 Proofs for the Policy Gradient Algorithm

This section proves a few key lemmas that are modified from [Tadić and Doucet, 2017] which leads to the convergence of the policy gradient algorithm analyzed in Section 4.3.2.

Let $\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}:=\boldsymbol{Q}_{\boldsymbol{\theta}}-\mathbf{1}\boldsymbol{v}_{\boldsymbol{\theta}}^\top$ and denote $\tilde{Q}_{\boldsymbol{\theta}}^t((s,a);(s',a'))$ to be the $((s,a),(s',a'))$th element of the tth power of $\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^t$. Under H4.11, we observe that $\|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^t\|\le\rho^tK_R$ for any $t\ge0$. For $i=1,...,d$, we also define the (s,a)th element of the $|\mathcal{S}||\mathcal{A}|$-dimensional gradient vector $\nabla_i\boldsymbol{\Pi}_{\boldsymbol{\theta}}$, and reward vector $\boldsymbol{r}$, respectively as:

$$\nabla_i\boldsymbol{\Pi}_{\boldsymbol{\theta}}(s,a):=\frac{\partial\log\Pi(a;s,\boldsymbol{\theta})}{\partial\eta_i},\quad r(s,a):=\mathcal{R}(s,a). \tag{4.7.1}$$

Using the above notations, the mean field in (4.3.27) can be evaluated as

$$h(\boldsymbol{\theta}) = \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \lambda^t \mathcal{R}(s',a') \tilde{Q}^t_{\boldsymbol{\theta}}((s,a);(s',a')) \nabla \log \Pi(a;s,\boldsymbol{\theta}) v_{\boldsymbol{\theta}}(s,a). \tag{4.7.2}$$

In particular, its ith element can be expressed as

$$h_i(\boldsymbol{\theta}) = \sum_{t=0}^{\infty} \lambda^t \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \tilde{\boldsymbol{Q}}^t_{\boldsymbol{\theta}} \boldsymbol{r} \ . \tag{4.7.3}$$

We also define the difference between $h(\boldsymbol{\theta})$ and $\nabla J(\boldsymbol{\theta})$ as

$$\Delta(\boldsymbol{\theta}) := h(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}). \tag{4.7.4}$$

4.7.1 Useful Lemmas

Lemma 5 *Let H4.10, H4.11 hold. For any $(\boldsymbol{\theta},\boldsymbol{\theta}') \in \Theta^2$ and $t \geq 0$, one has*

$$\|\boldsymbol{Q}^t_{\boldsymbol{\theta}} - \boldsymbol{Q}^t_{\boldsymbol{\theta}'}\| \leq C_1 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|, \quad \|\tilde{\boldsymbol{Q}}^t_{\boldsymbol{\theta}} - \tilde{\boldsymbol{Q}}^t_{\boldsymbol{\theta}'}\| \leq C_1 (t\rho^t) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \ , \tag{4.7.5}$$

where we have set $C_1 := \rho K_R^2 (2\bar{b} + L_Q) + L_Q$ in the above.

Proof For part 1), we observe that each entry of $\boldsymbol{Q}_{\boldsymbol{\theta}}$ is given by [cf. (4.3.19)]:

$$Q_{\boldsymbol{\theta}}((s,a);(s',a')) := \Pi(a';s',\boldsymbol{\theta}) P^a_{s,s'} \ ,$$

which is Lipschitz continuous *w.r.t.* $\boldsymbol{\theta}$ since

$$\nabla \Pi(a|s,\boldsymbol{\theta}) = \\ -\Big(\sum_{a'\in\mathcal{A}} \exp\big(\langle \boldsymbol{\theta} \,|\, \boldsymbol{x}(s,a') - \boldsymbol{x}(s,a)\rangle\big)\Big)^{-2} \sum_{a'\in\mathcal{A}} \exp\big(\langle \boldsymbol{\theta} \,|\, \boldsymbol{x}(s,a') - \boldsymbol{x}(s,a)\rangle\big)(\boldsymbol{x}(s,a') - \boldsymbol{x}(s,a))$$

is bounded by $\max_{s,a,a'} \|\boldsymbol{x}(s,a') - \boldsymbol{x}(s,a)\| \leq 2\bar{b}$ [cf. H4.10]. This implies

$$|Q_{\boldsymbol{\theta}}((s,a);(s',a')) - Q_{\boldsymbol{\theta}'}((s,a);(s',a'))| \leq 2\bar{b} |P^a_{s,s'}| \, \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \ . \tag{4.7.6}$$

Since $|P^a_{s,s'}| \leq 1$ for any s, s', a, we have $\|\boldsymbol{Q}_{\boldsymbol{\theta}} - \boldsymbol{Q}_{\boldsymbol{\theta}'}\| \leq 2\bar{b}\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

For any $\boldsymbol{\theta} \in \Theta$ and any $t \geq 0$, we have:

$$\begin{aligned} \tilde{\boldsymbol{Q}}^{t+1}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{Q}}^{t+1}_{\boldsymbol{\theta}'} &= \sum_{\tau=0}^{t} \tilde{\boldsymbol{Q}}^{\tau}_{\boldsymbol{\theta}} (\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}'}) \tilde{\boldsymbol{Q}}^{t-\tau}_{\boldsymbol{\theta}'} \\ &= \sum_{\tau=0}^{t} \tilde{\boldsymbol{Q}}^{\tau}_{\boldsymbol{\theta}} (\boldsymbol{Q}_{\boldsymbol{\theta}} - \boldsymbol{Q}_{\boldsymbol{\theta}'} - \mathbf{1}(\boldsymbol{v}_{\boldsymbol{\theta}} - \boldsymbol{v}_{\boldsymbol{\theta}'})^\top) \tilde{\boldsymbol{Q}}^{t-\tau}_{\boldsymbol{\theta}'} \ . \end{aligned} \tag{4.7.7}$$

As such,

$$\begin{aligned}\|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^{t+1} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}'}^{t+1}\| &\leq \sum_{\tau=0}^{t} \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^{\tau}\| \|\boldsymbol{Q}_{\boldsymbol{\theta}} - \boldsymbol{Q}_{\boldsymbol{\theta}'} - \mathbf{1}(\boldsymbol{v}_{\boldsymbol{\theta}} - \boldsymbol{v}_{\boldsymbol{\theta}'})^{\top}\| \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}'}^{t-\tau}\| \\ &\leq K_R^2 \sum_{\tau=0}^{t} \rho^{\tau} \rho^{t-\tau} (\|\boldsymbol{Q}_{\boldsymbol{\theta}} - \boldsymbol{Q}_{\boldsymbol{\theta}'}\| + \|\boldsymbol{v}_{\boldsymbol{\theta}} - \boldsymbol{v}_{\boldsymbol{\theta}'}\|) \\ &\leq K_R^2 (2\bar{b} + L_Q)(t\,\rho^t) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,.\end{aligned} \tag{4.7.8}$$

Consequently,

$$\begin{aligned}\|\boldsymbol{Q}_{\boldsymbol{\theta}}^{t+1} - \boldsymbol{Q}_{\boldsymbol{\theta}'}^{t+1}\| &\leq \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^{t+1} - \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}'}^{t+1}\| + \|\boldsymbol{v}_{\boldsymbol{\theta}} - \boldsymbol{v}_{\boldsymbol{\theta}'}\| \\ &\leq (K_R^2 (t\,\rho^t)(2\bar{b} + L_Q) + L_Q) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,.\end{aligned} \tag{4.7.9}$$

Setting $C_1 = \rho K_R^2 (2\bar{b} + L_Q) + L_Q$ completes the proof. ■

Lemma 6 *Let H4.10, H4.11 hold. The following statements are true:*

1. *The average reward $J(\boldsymbol{\theta})$ is differentiable and for any $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, one has*

$$\|\nabla J(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}')\| \leq \mathrm{R}_{\max}\, |\mathcal{S}||\mathcal{A}|\, L_v \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,. \tag{4.7.10}$$

2. *For any $\boldsymbol{\theta} \in \Theta$, one has*

$$\|\Delta(\boldsymbol{\theta})\| \leq 2\bar{b}\, \mathrm{R}_{\max}\, K_R \frac{1-\lambda}{(1-\rho)^2} \,. \tag{4.7.11}$$

Proof For part 1), we observe that

$$J(\boldsymbol{\theta}) = \mathbb{E}_{(S,A)\sim \boldsymbol{v}_{\boldsymbol{\theta}}}[\mathcal{R}(S,A)] = \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} v_{\boldsymbol{\theta}}(s,a)\mathcal{R}(s,a) \,. \tag{4.7.12}$$

It follows from the Lipschitz continuity of $\mathrm{J}_{\boldsymbol{v}_{\boldsymbol{\theta}}}^{\boldsymbol{\theta}}(\boldsymbol{\theta})$ [cf. H4.11] that

$$\begin{aligned}\|\nabla J(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}')\| &\leq \sum_{(s,a)\in\mathcal{S}\times\mathcal{A}} |\mathcal{R}(s,a)| \|\nabla v_{\boldsymbol{\theta}}(s,a) - \nabla v_{\boldsymbol{\theta}'}(s,a)\| \\ &\leq \mathrm{R}_{\max}\, |\mathcal{S}||\mathcal{A}|\, L_v \,\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,.\end{aligned} \tag{4.7.13}$$

The above verifies (4.7.10).

For part 2), we define

$$J_T(\boldsymbol{\theta}, (s,a)) := \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s',a') Q_{\boldsymbol{\theta}}^T((s,a);(s',a')) \,, \tag{4.7.14}$$

$$g(\boldsymbol{\theta}) := \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s,a)\tilde{Q}_{\boldsymbol{\theta}}^t((s,a);(s',a'))\nabla \log \Pi(a;s,\boldsymbol{\theta})v_{\boldsymbol{\theta}}(s,a) \; . \tag{4.7.15}$$

As shown in [Tadić and Doucet, 2017, Lemma 8.2], we have $\lim_{T\to\infty} \nabla_{\boldsymbol{\theta}} J_T(\boldsymbol{\theta},(s,a)) = g(\boldsymbol{\theta})$ for all $\boldsymbol{\theta}\in\Theta$ and $(s,a)\in\mathcal{S}\times\mathcal{A}$. As such

$$\begin{aligned}
&\Delta(\boldsymbol{\theta}) = h(\boldsymbol{\theta}) - g(\boldsymbol{\theta}) \\
&= \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} (\lambda^t - 1)\mathcal{R}(s,a)\tilde{Q}_{\boldsymbol{\theta}}^t((s,a);(s',a'))\nabla \log \Pi(a;s,\boldsymbol{\theta})v_{\boldsymbol{\theta}}(s,a) \; .
\end{aligned} \tag{4.7.16}$$

and in particular, the ith element is given by

$$\Delta_i(\boldsymbol{\theta}) = \sum_{t=0}^{\infty} \sum_{(s,a),(s',a')\in\mathcal{S}\times\mathcal{A}} (\lambda^t - 1)\boldsymbol{v}_{\boldsymbol{\theta}}^{\top} \mathrm{Diag}(\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\theta}})\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^t \boldsymbol{r} \; , \tag{4.7.17}$$

which can be bounded as

$$\begin{aligned}
|\Delta_i(\boldsymbol{\theta})| &\leq \sum_{t=0}^{\infty} (1-\lambda^t)\|\boldsymbol{v}_{\boldsymbol{\theta}}\|\|\nabla_i \boldsymbol{\Pi}_{\boldsymbol{\theta}}\|_{\infty}\|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^t\|\|\boldsymbol{r}\| \\
&\overset{(a)}{\leq} 2\bar{b}\; \mathrm{R}_{\max}\, K_R \sum_{t=0}^{\infty}(1-\lambda^t)\rho^t \leq 2\bar{b}\; \mathrm{R}_{\max}\, K_R \frac{1-\lambda}{(1-\rho)^2} \; ,
\end{aligned} \tag{4.7.18}$$

where (a) uses H4.11, H4.10, and Proposition 9. The above implies that $\|\Delta(\boldsymbol{\theta})\| \leq 2\bar{b}\; \mathrm{R}_{\max}\, K_R \frac{1-\lambda}{(1-\rho)^2}$. ■

Lemma 7 *Let H4.10, H4.11 hold. Denote the joint state x as $x = (s,a,g) \in \mathcal{S}\times\mathcal{A}\times\mathbb{R}^d$. There exists $\delta\in[0,1)$, $C_2\in[1,\infty)$ such that for any $t\geq 0$,*

$$\begin{aligned}
&\|P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})\| \leq C_2\, t\, \delta^t(1+\|g\|) \; , \\
&\left\|(P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})) - (P_{\boldsymbol{\theta}'}^t H_{\boldsymbol{\theta}'}(x) - h(\boldsymbol{\theta}'))\right\| \leq C_2\, t\, \delta^t\, \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|(1+\|g\|) \; .
\end{aligned} \tag{4.7.19}$$

Proof Denote the joint state as $x=(s,a,g)$, we observe that

$$\begin{aligned}
&P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\theta}}(x) = \mathbb{E}_{\Pi_{\boldsymbol{\theta}}}[\mathcal{R}(S_t,A_t)G_t \,|\, (S_0,A_0)=(s,a), G_0 = g] \\
&= \mathbb{E}_{\Pi_{\boldsymbol{\theta}}}\left[\mathcal{R}(S_t,A_t)\Big(\lambda^t g + \sum_{i=1}^{t-1}\lambda^i \nabla \log \Pi(A_i;S_i,\boldsymbol{\theta})\Big) \,|\, (S_0,A_0)=(s,a)\right] \\
&= \sum_{i=0}^{t-1} \sum_{(s',a'),(s'',a'')\in\mathcal{S}\times\mathcal{A}} \lambda^i \mathcal{R}(s'',a'')Q_{\boldsymbol{\theta}}^i((s',a');(s'',a''))\nabla \log \Pi(a';s',\boldsymbol{\theta})Q_{\boldsymbol{\theta}}^{t-i}((s,a);(s',a')) \\
&\quad + \lambda^t g \sum_{(s',a')\in\mathcal{S}\times\mathcal{A}} \mathcal{R}(s',a')Q_{\boldsymbol{\theta}}^t((s,a);(s',a')) \; .
\end{aligned}$$

The jth element of the above is thus given by

$$\big[P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\theta}}(x)\big]_j = \sum_{i=0}^{t-i} \lambda^i \boldsymbol{e}_{(s,a)}^\top \boldsymbol{Q}_{\boldsymbol{\theta}}^{t-i} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r} + \lambda^t g_j \mathbf{1}^\top \boldsymbol{Q}_{\boldsymbol{\theta}}^t \boldsymbol{r} \,, \tag{4.7.20}$$

where g_j is the jth element of g and $\boldsymbol{e}_{(s,a)}$ is the (s,a)th coordinate vector. Moreover, we recall that

$$h_j(\boldsymbol{\theta}) = \sum_{t=0}^{\infty} \lambda^t \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^t \boldsymbol{r} \,. \tag{4.7.21}$$

Note that

$$\begin{aligned} \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \mathbf{1} &= \sum_{(s,a) \in \mathcal{S} \times \mathcal{A}} v_{\boldsymbol{\theta}}(s,a) \nabla_j \log \Pi(a; s, \boldsymbol{\theta}) \\ &= \sum_{s \in \mathcal{S}} \Big(\sum_{a \in \mathcal{A}} \underbrace{\Pi(a; s, \boldsymbol{\theta}) \nabla_j \log \Pi(a; s, \boldsymbol{\theta})}_{= \nabla_j \Pi(a; s, \boldsymbol{\theta})} \Big) \overline{\Pi}_{\boldsymbol{\theta}}(s) = 0 \,. \end{aligned} \tag{4.7.22}$$

where we recalled that $\overline{\Pi}_{\boldsymbol{\theta}}(s)$ is the stationary distribution for the MDP on the state. Using the decomposition $\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^t = \boldsymbol{Q}_{\boldsymbol{\theta}}^t - \mathbf{1} \boldsymbol{v}_{\boldsymbol{\theta}}^\top$, we observe

$$\begin{aligned} h_j(\boldsymbol{\theta}) &= \sum_{i=0}^{t-1} \lambda^i \Big\{ \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r} - \underbrace{\boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \mathbf{1}}_{=0} \boldsymbol{v}_{\boldsymbol{\theta}}^\top \boldsymbol{r} \Big\} + \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^i \boldsymbol{r} \\ &= \sum_{i=0}^{t-1} \lambda^i \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r} + \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^i \boldsymbol{r} \,. \end{aligned}$$

Therefore,

$$\begin{aligned} &\big[P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\theta}}(x)\big]_j - h_j(\boldsymbol{\theta}) \\ &= \sum_{i=0}^{t-1} \lambda^i \Big\{ \boldsymbol{e}_{(s,a)}^\top (\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^{t-i} + \mathbf{1} \boldsymbol{v}_{\boldsymbol{\theta}}) \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r} - \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r} \Big\} \\ &\quad + \lambda^t g_j \mathbf{1}^\top \boldsymbol{Q}_{\boldsymbol{\theta}}^t \boldsymbol{r} - \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^i \boldsymbol{r} \\ &= \sum_{i=0}^{t-1} \lambda^i \boldsymbol{e}_{(s,a)}^\top \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^{t-i} \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r} + \lambda^t g_j \mathbf{1}^\top \boldsymbol{Q}_{\boldsymbol{\theta}}^t \boldsymbol{r} - \sum_{i=t}^{\infty} \lambda^i \boldsymbol{v}_{\boldsymbol{\theta}}^\top \mathrm{Diag}(\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}) \tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^i \boldsymbol{r} \,. \end{aligned} \tag{4.7.23}$$

Consequently, we obtain the upper bound as

$$\begin{aligned} \big|\big[P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\theta}}(x)\big]_j - h_j(\boldsymbol{\theta})\big| \leq &\sum_{i=0}^{t-1} \lambda^i \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^{t-i}\| \|\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}\|_\infty \|\boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r}\| + \lambda^t |g_j| \|\boldsymbol{Q}_{\boldsymbol{\theta}}^t \boldsymbol{r}\| \\ &+ \sum_{i=t}^{\infty} \lambda^i \|\boldsymbol{v}_{\boldsymbol{\theta}}\| \|\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}\|_\infty \|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^i \boldsymbol{r}\| \,. \end{aligned} \tag{4.7.24}$$

Using H4.10, H4.11 and notice that $\|\nabla_j \boldsymbol{\Pi}_{\boldsymbol{\theta}}\|_\infty \leq 2\overline{b}$, $\|\boldsymbol{Q}_{\boldsymbol{\theta}}^i \boldsymbol{r}\| \leq \overline{R}$, $\|\tilde{\boldsymbol{Q}}_{\boldsymbol{\theta}}^i \boldsymbol{r}\| \leq$

$\overline{R}K_R\sqrt{|\mathcal{S}||\mathcal{A}|}\rho^i$, we obtain

$$\left|\left[P^t_{\boldsymbol{\eta}}H_{\boldsymbol{\theta}}(x)\right]_j - h_j(\boldsymbol{\theta})\right| \le 2\bar{b}\,\overline{R}\,K_R\sum_{i=0}^{t-1}\lambda^i\rho^{t-i} + \lambda^t|g_j|\overline{R} + 2\bar{b}\,\overline{R}K_R\sqrt{|\mathcal{S}||\mathcal{A}|}\sum_{i=t}^{\infty}\lambda^i\rho^i\ . \tag{4.7.25}$$

Observe that each of the above term decays geometrically with t, as such there exists $C_2' \in [1,\infty)$, $\delta \in [0,1)$ such that [1]

$$\left|\left[P^t_{\boldsymbol{\eta}}H_{\boldsymbol{\theta}}(x)\right]_j - h_j(\boldsymbol{\theta})\right| \le C_2'(t\delta^t)(1+\|g\|)\ , \tag{4.7.26}$$

which naturally implies the first equation in (4.7.19).

For the second equation in (4.7.19),

$$\begin{aligned}
&\left[P^t_{\boldsymbol{\eta}}H_{\boldsymbol{\theta}}(x)\right]_j - h_j(\boldsymbol{\theta}) - \left\{\left[P^t_{\boldsymbol{\theta}'}H_{\boldsymbol{\theta}'}(x)\right]_j - h_j(\boldsymbol{\theta}')\right\}\\
&= \sum_{i=0}^{t-1}\lambda^i \boldsymbol{e}_{(s,a)}^\top\{\tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\theta}}\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}})\boldsymbol{Q}^i_{\boldsymbol{\theta}} - \tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\theta}'}\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}'})\boldsymbol{Q}^i_{\boldsymbol{\theta}'}\}\boldsymbol{r}\\
&\quad + \lambda^t g_j\mathbf{1}^\top(\boldsymbol{Q}^t_{\boldsymbol{\theta}} - \boldsymbol{Q}^t_{\boldsymbol{\theta}'})\boldsymbol{r} + \sum_{i=t}^{\infty}\lambda^i\{\boldsymbol{v}_{\boldsymbol{\theta}'}^\top\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}'})\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\theta}'} - \boldsymbol{v}_{\boldsymbol{\theta}}^\top\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}})\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\theta}}\}\boldsymbol{r}\ .
\end{aligned} \tag{4.7.27}$$

This leads to the upper bound:

$$\begin{aligned}
&\left|\left[P^t_{\boldsymbol{\eta}}H_{\boldsymbol{\theta}}(x)\right]_j - h_j(\boldsymbol{\theta}) - \left\{\left[P^t_{\boldsymbol{\theta}'}H_{\boldsymbol{\theta}'}(x)\right]_j - h_j(\boldsymbol{\theta}')\right\}\right|\\
&\le \sqrt{|\mathcal{S}||\mathcal{A}|}\overline{R}\sum_{i=0}^{t-i}\lambda^i\|\tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\theta}}\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}})\boldsymbol{Q}^i_{\boldsymbol{\theta}} - \tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\theta}'}\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}'})\boldsymbol{Q}^i_{\boldsymbol{\theta}'}\|\\
&\quad + \lambda^t|\mathcal{S}||\mathcal{A}|\|\boldsymbol{Q}^t_{\boldsymbol{\theta}} - \boldsymbol{Q}^t_{\boldsymbol{\theta}'}\| + \sqrt{|\mathcal{S}||\mathcal{A}|}\overline{R}\sum_{i=t}^{\infty}\lambda^i\|\boldsymbol{v}_{\boldsymbol{\theta}'}^\top\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}'})\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\theta}'} - \boldsymbol{v}_{\boldsymbol{\theta}}^\top\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}})\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\theta}}\|\ .
\end{aligned} \tag{4.7.28}$$

Using the boundedness and Lipschitz continuity of $\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}}$, $\boldsymbol{v}_{\boldsymbol{\theta}}$, $\boldsymbol{Q}^t_{\boldsymbol{\theta}}$, $\tilde{\boldsymbol{Q}}^t_{\boldsymbol{\theta}}$ [cf. Lemma 5], let $C_{2,1}, C_{2,2} \in [1,\infty)$, the norms in the above can be bounded as

$$\begin{aligned}
&\|\tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\theta}}\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}})\boldsymbol{Q}^i_{\boldsymbol{\theta}} - \tilde{\boldsymbol{Q}}^{t-i}_{\boldsymbol{\theta}'}\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}'})\boldsymbol{Q}^i_{\boldsymbol{\theta}'}\| \le C_{2,1}\big((t-i)\rho^{t-i}\big)\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\\
&\|\boldsymbol{v}_{\boldsymbol{\theta}'}^\top\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}'})\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\theta}'} - \boldsymbol{v}_{\boldsymbol{\theta}}^\top\mathrm{Diag}(\nabla_j\boldsymbol{\Pi}_{\boldsymbol{\theta}})\tilde{\boldsymbol{Q}}^i_{\boldsymbol{\theta}}\| \le C_{2,2}(i\rho^i)\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\\
&\|\boldsymbol{Q}^t_{\boldsymbol{\theta}} - \boldsymbol{Q}^t_{\boldsymbol{\theta}'}\| \le C_1\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\ .
\end{aligned} \tag{4.7.29}$$

The above shows that the three terms in the right hand side of (4.7.28) are proportional to $(1+\|g\|)\|\boldsymbol{\theta}-\boldsymbol{\theta}'\|$ and decay geometrically with t. This implies there exists $C_2'' \in [1,\infty)$,

1. Note that an exact characterization for C_2' is also possible.

$\delta \in [0,1)$ such that

$$\left\| P^t_{\boldsymbol{\eta}} H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta}) - \left\{ P^t_{\boldsymbol{\theta}'} H_{\boldsymbol{\theta}'}(x) - h(\boldsymbol{\theta}') \right\} \right\| \leq C_2''(t\delta^t)(1 + \|g\|)\|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,. \tag{4.7.30}$$

Setting $C_2 = \max\{C_2', C_2''\}$ concludes the proof of the current lemma. ■

4.7.2 Proof of Proposition 9

Proposition *Under H4.10, it holds for any* $(\boldsymbol{\theta}, \boldsymbol{\theta}') \in \Theta^2$, $(s,a) \in \mathsf{S} \times \mathsf{A}$,

$$\|\nabla \log \Pi_{\boldsymbol{\theta}}(a;s)\| \leq 2\bar{b}, \quad \|\nabla \log \Pi_{\boldsymbol{\theta}}(a;s) - \nabla \log \Pi_{\boldsymbol{\theta}'}(a;s)\| \leq 8\bar{b}^2 \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,. \tag{4.7.31}$$

Proof To simplify notations, let us define $\Delta \boldsymbol{x}(a,b) := \boldsymbol{x}(s,a) - \boldsymbol{x}(s,b)$ as the difference between two features. The proof is straightforward as we observe that

$$\nabla \log \Pi_{\boldsymbol{\theta}}(a;s) = \frac{1}{\sum_{a' \in \mathsf{A}} \exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle \right)} \sum_{b \in \mathsf{A}} \exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle \right) \Delta \boldsymbol{x}(a,b) \,. \tag{4.7.32}$$

Observe that

$$\|\nabla \log \Pi_{\boldsymbol{\theta}}(a;s)\| \leq \max_{a,b \in \mathsf{A}} \|\boldsymbol{x}(s,a) - \boldsymbol{x}(s,b)\| \leq 2\bar{b} \,. \tag{4.7.33}$$

Moreover, the Hessian of the log policy can be evaluated as:

$$\begin{aligned} &\nabla^2 \log \Pi_{\boldsymbol{\theta}}(a;s) = \\ &\frac{1}{\sum_{a' \in \mathsf{A}} \exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle \right)} \sum_{b \in \mathsf{A}} \exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle \right) \Delta \boldsymbol{x}(a,b) \Delta \boldsymbol{x}(b,a)^\top - \\ &\Big(\sum_{b \in \mathsf{A}} \frac{\exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle \right)}{\sum_{a' \in \mathsf{A}} \exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle \right)} \Delta \boldsymbol{x}(a,b) \Big) \Big(\frac{\exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(b,a) \rangle \right)}{\sum_{a' \in \mathsf{A}} \exp\left(\langle \boldsymbol{\theta} \,|\, \Delta \boldsymbol{x}(a',a) \rangle \right)} \Delta \boldsymbol{x}(a,b) \Big)^\top \,. \end{aligned} \tag{4.7.34}$$

It can be checked that

$$\|\nabla^2 \log \Pi_{\boldsymbol{\theta}}(a;s)\| \leq \max_{a,b \in \mathsf{A}} \|\Delta \boldsymbol{x}(a,b) \Delta \boldsymbol{x}(b,a)^\top\| + \left(\max_{a,b \in \mathsf{A}} \|\Delta \boldsymbol{x}(a,b)\| \right)^2 \leq 8\bar{b}^2 \,. \tag{4.7.35}$$

This implies smoothness condition in (4.3.28). ■

4.7.3 Proof of Proposition 10

Proposition *Under H4.10, H4.11, the function*

$$\hat{H}_{\boldsymbol{\eta}}(x) = \textstyle\sum_{t=0}^{\infty} \left\{ P^t_{\boldsymbol{\eta}} H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}) \right\} \,, \tag{4.7.36}$$

is well defined and satisfies the Poisson equation (4.2.4). For all $x \in \mathsf{X}$, $(\boldsymbol{\eta}, \boldsymbol{\eta}') \in \Theta^2$, there exists constants $L_{PH}^{(0)}$, $L_{PH}^{(1)}$ such that

$$\max\{\|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x)\|, \|\hat{H}_{\boldsymbol{\eta}}(x)\|\} \le L_{PH}^{(0)}, \quad \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'}\hat{H}_{\boldsymbol{\eta}'}(x)\| \le L_{PH}^{(1)}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \,. \quad (4.7.37)$$

Moreover, the constants are in the order of $L_{PH}^{(0)} = \mathcal{O}(\frac{1}{1-\max\{\rho,\lambda\}})$, $L_{PH}^{(1)} = \mathcal{O}(\frac{1}{1-\max\{\rho,\lambda\}})$.

Proof From Lemma 7, there exists $C_2 \in [1, \infty)$, $\delta \in [0, 1)$ such that

$$\|P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\| \le C_2\, t\, \delta^t (1 + \|g\|), \ \forall\, t \ge 1, \ \forall\, x \in \mathsf{X} \,, \quad (4.7.38)$$

where we recall that $\delta = \max\{\rho, \lambda\}$. It follows that the solution to the Poisson equation $\hat{H}_{\boldsymbol{\eta}}(x)$ in (4.3.30) is well defined.

Moreover, it satisfies (4.2.4) and

$$\max\{\|\hat{H}_{\boldsymbol{\eta}}(x)\|, \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x)\|\} \le L_{PH}^{(0)} \,, \quad (4.7.39)$$

for some $L_{PH}^{(0)} = \mathcal{O}(\frac{1}{1-\max\{\rho,\lambda\}}) < \infty$ (note that g is bounded as specified by the state space X). As such, the first equation in (4.3.31) of the proposition is proven. Finally, applying the definition of $\hat{H}_{\boldsymbol{\eta}}(x)$ shows that

$$P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}'}(x) = \sum_{t=1}^{\infty} \Big\{ (P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})) - (P_{\boldsymbol{\eta}'}^t H_{\boldsymbol{\eta}'}(x) - h(\boldsymbol{\eta}')) \Big\} \,. \quad (4.7.40)$$

Using Lemma 7, this implies

$$\begin{aligned} \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}'}(x)\| &\le \sum_{t=1}^{\infty} \Big\| (P_{\boldsymbol{\eta}}^t H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})) - (P_{\boldsymbol{\eta}'}^t H_{\boldsymbol{\eta}'}(x) - h(\boldsymbol{\eta}')) \Big\| \\ &\le \sum_{t=1}^{\infty} \Big\{ C_2 (t\delta^t)(1 + \|g\|)\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \Big\} \,. \end{aligned} \quad (4.7.41)$$

As such, there exists $L_{PH}^{(1)} = \mathcal{O}(\frac{1}{1-\max\{\rho,\lambda\}}) \in [1, \infty)$ such that

$$\|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}'}(x)\| \le L_{PH}^{(1)}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\| \,, \quad (4.7.42)$$

for all $x \in \mathsf{X}$. This proves the second equation in (4.3.31) of the proposition. ■

4.7.4 Proof of Proposition 11

Proposition *Under H4.10, H4.11, the gradient $\nabla J(\boldsymbol{\theta})$ is $\mathrm{R}_{\max}\,|\mathcal{S}||\mathcal{A}|$-Lipschitz continuous. Moreover, for any $\boldsymbol{\theta} \in \Theta$, it holds that*

$$(1-\lambda)^2\Gamma^2 + 2\,\langle \nabla J(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta})\rangle \ge \|h(\boldsymbol{\theta})\|^2, \ \|\nabla J(\boldsymbol{\theta})\| \le \|h(\boldsymbol{\theta})\| + (1-\lambda)\Gamma \,, \quad (4.7.43)$$

where $\Gamma := 2\bar{b}\ \mathrm{R}_{\max} K_R \frac{1}{(1-\rho)^2}$.

Proof The first statement is a direct application of part 1) in Lemma 6 which holds under H4.10, H4.11. To prove the second statement, let us define the error vector as

$$\Delta(\boldsymbol{\theta}) := h(\boldsymbol{\theta}) - \nabla J(\boldsymbol{\theta}) \tag{4.7.44}$$

Applying Lemma 6 shows that $\sup_{\boldsymbol{\theta}\in\Theta} \|\Delta(\boldsymbol{\theta})\|^2 \leq \Gamma^2(1-\lambda)^2$. We observe that

$$\begin{aligned} \langle \nabla J(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle &= \langle h(\boldsymbol{\theta}) - \Delta(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle = \|h(\boldsymbol{\theta})\|^2 - \langle \Delta(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle \\ &\geq \|h(\boldsymbol{\theta})\|^2 - \frac{1}{2}(\|h(\boldsymbol{\theta})\|^2 + \|\Delta(\boldsymbol{\theta})\|^2) \ . \end{aligned} \tag{4.7.45}$$

This implies

$$\frac{\Gamma^2}{2}(1-\lambda)^2 + \langle \nabla J(\boldsymbol{\theta}) \,|\, h(\boldsymbol{\theta}) \rangle \geq \frac{1}{2}\|h(\boldsymbol{\theta})\|^2 \ . \tag{4.7.46}$$

Furthermore, it is straightforward to show that

$$\|\nabla J(\boldsymbol{\theta})\| \leq \|h(\boldsymbol{\theta})\| + \|\Delta(\boldsymbol{\theta})\| \leq \|h(\boldsymbol{\theta})\| + \Gamma(1-\lambda) \ , \tag{4.7.47}$$

which concludes the proof. ■

4.8 Existence and Regularity of the Solutions of Poisson Equations

Consider the following assumptions:

H4.12 *For any* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, *we have* $\sup_{x\in\mathsf{X}} \|P_{\boldsymbol{\eta}}(x,\cdot) - P_{\boldsymbol{\theta}'}(x,\cdot)\|_{\mathrm{TV}} \leq L_P\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

H4.13 *For any* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, *we have* $\sup_{x\in\mathsf{X}} \|H_{\boldsymbol{\theta}}(x) - H_{\boldsymbol{\theta}'}(x)\| \leq L_H\|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

H4.14 *There exists* $\rho < 1$, $K_P < \infty$ *such that*

$$\sup_{\boldsymbol{\theta}\in\mathbb{R}^d, x\in\mathsf{X}} \|P^n_{\boldsymbol{\eta}}(x,\cdot) - \pi_{\boldsymbol{\theta}}(\cdot)\|_{\mathrm{TV}} \leq \rho^n K_P, \tag{4.8.1}$$

Lemma 8 *Assume H4.12–4.14. Then, for any* $\boldsymbol{\theta} \in \Theta$ *and* $x \in \mathsf{X}$,

$$\|\hat{H}_{\boldsymbol{\theta}}(x)\| \leq \frac{\sigma K_P}{1-\rho} \ , \tag{4.8.2}$$

$$\|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\theta}}(x)\| \leq \frac{\sigma\rho K_P}{1-\rho} \ . \tag{4.8.3}$$

Moreover, for $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$ *and* $x \in \mathsf{X}$,

$$\left\| P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\theta}}(x) - P_{\boldsymbol{\theta}'} \hat{H}_{\boldsymbol{\theta}'}(x) \right\| \leq L_{PH}^{(1)} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,, \tag{4.8.4}$$

where

$$L_{PH}^{(1)} = \frac{K_P^2 \sigma L_P}{(1-\rho)^2}(2 + K_P) + \frac{K_P}{1-\rho} L_H \,. \tag{4.8.5}$$

Proof Note that, under H4.14,

$$\begin{aligned} &\sum_{i=0}^{\infty} \left\| P_{\boldsymbol{\eta}}^i (H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})) - \pi_{\boldsymbol{\theta}}(H_{\boldsymbol{\theta}}(\cdot) - h(\boldsymbol{\theta})) \right\| \\ &\leq \|H_{\boldsymbol{\theta}}(\cdot) - h(\boldsymbol{\theta})\|_\infty \, K_P \sum_{i=0}^{\infty} \rho^i \leq \frac{\sigma K_P}{1-\rho} \,. \end{aligned} \tag{4.8.6}$$

Therefore, for all $\boldsymbol{\theta} \in \Theta$ and $x \in \mathsf{X}$, the series

$$\sum_{i=0}^{\infty} P_{\boldsymbol{\eta}}^i (H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})) - \pi_{\boldsymbol{\theta}}(H_{\boldsymbol{\theta}}(\cdot) - h(\boldsymbol{\theta})) \tag{4.8.7}$$

is uniformly converging and is a solution of the Poisson equation (4.2.4). In addition, (4.8.2) and (4.8.3) follow directly from (4.8.6). Under H4.14, applying a simple modification [2] of [Fort et al., 2011, Lemma 4.2, 1st statement] shows [3] that for any $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, we have

$$\|\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_{\mathrm{TV}} \leq \frac{K_P(1+K_P)}{1-\rho} \sup_{x \in \mathsf{X}} \|P_{\boldsymbol{\eta}}(x, \cdot) - P_{\boldsymbol{\theta}'}(x, \cdot)\|_{\mathrm{TV}} \,. \tag{4.8.8}$$

Again using a simple modification of [Fort et al., 2011, Lemma 4.2, 2nd statement] shows that for any $X \in \mathsf{X}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathbb{R}^d$, it holds

$$\begin{aligned} &\left\| P_{\boldsymbol{\eta}} \hat{H}_{\boldsymbol{\theta}}(x) - P_{\boldsymbol{\theta}'} \hat{H}_{\boldsymbol{\theta}'}(x) \right\| \\ &\leq \frac{K_P^2}{(1-\rho)^2} \Big(\sup_{\boldsymbol{\theta} \in \Theta, x \in \mathsf{X}} \|H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})\| \Big) \Big(\sup_{x \in \mathsf{X}} \|P_{\boldsymbol{\eta}}(x, \cdot) - P_{\boldsymbol{\theta}'}(x, \dot{o}t)\|_{\mathrm{TV}} \Big) \\ &\quad + \frac{K_P}{1-\rho} \Big(\sup_{\boldsymbol{\theta} \in \Theta, x \in \mathsf{X}} \|H_{\boldsymbol{\theta}}(x) - h(\boldsymbol{\theta})\| \Big) \|\pi_{\boldsymbol{\theta}} - \pi_{\boldsymbol{\theta}'}\|_{\mathrm{TV}} + \frac{K_P}{1-\rho} \sup_{x \in \mathsf{X}} \|H_{\boldsymbol{\theta}}(x) - H_{\boldsymbol{\theta}'}(x)\| \\ &\leq \left(\frac{K_P^2 \sigma L_P}{(1-\rho)^2}(2 + K_P) + \frac{K_P}{1-\rho} L_H \right) \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| = L_{PH}^{(1)} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| \,, \end{aligned} \tag{4.8.9}$$

where the last inequality is due to H4.12, H4.13, H4.7 and (4.8.8). ■

2. We note that under H4.14, the constants $\rho_\theta, \rho_{\theta'}$ are the same in [Fort et al., 2011, Lemma 4.2] which simplifies the derivation and yields a tighter bound.

3. Note that we take the measurable function as $V = 1$ therein.

Part II

FAST MAXIMUM LIKELIHOOD ESTIMATION

Chapter 5

Fast Incremental EM Methods

Abstract: *The EM algorithm is one of the most popular algorithm for inference in latent data models. The original formulation of the EM algorithm does not scale to large data set, because the whole data set is required at each iteration of the algorithm. To alleviate this problem, Neal and Hinton [1998] have proposed an incremental version of the EM (iEM) in which at each iteration the conditional expectation of the latent data (E-step) is updated only for a mini-batch of observations. Another approach has been proposed by Cappé and Moulines [2009] in which the E-step is replaced by a stochastic approximation step, closely related to stochastic gradient. In this paper, we analyze incremental and stochastic version of the EM algorithm as well as the variance reduced-version of [Chen et al., 2018] in a common unifying framework. We also introduce a new version incremental version, inspired by the SAGA algorithm by Defazio et al. [2014]. We establish non-asymptotic convergence bounds for global convergence. Numerical applications are presented in this article to illustrate our findings. This chapter corresponds to the article [Karimi et al., 2019c].*

Contents

5.1 Introduction

Many problems in machine learning pertain to tackling an empirical risk minimization of the form

$$\min_{\boldsymbol{\theta}\in\Theta} \overline{\mathcal{L}}(\boldsymbol{\theta}) := \mathrm{R}(\boldsymbol{\theta}) + \mathcal{L}(\boldsymbol{\theta}) \quad \text{with} \quad \mathcal{L}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \{ -\log g(y_i;\boldsymbol{\theta})\} \,, \quad (5.1.1)$$

where $\{y_i\}_{i=1}^n$ are the observations, Θ is a convex subset of $\mathbb{R}^d$ for the parameters, $\mathrm{R} : \Theta \to \mathbb{R}$ is a smooth convex regularization function and for each $\boldsymbol{\theta} \in \Theta$, $g(y;\boldsymbol{\theta})$ is the (incomplete) likelihood of each individual observation. The objective function $\overline{\mathcal{L}}(\boldsymbol{\theta})$ is possibly *non-convex* and is assumed to be lower bounded $\overline{\mathcal{L}}(\boldsymbol{\theta}) > -\infty$ for all $\boldsymbol{\theta} \in \Theta$. In the latent variable model, $g_i(y_i;\boldsymbol{\theta})$, is the marginal of the complete data likelihood defined as $f_i(z_i, y_i;\boldsymbol{\theta})$, i.e. $g_i(y_i;\boldsymbol{\theta}) = \int_{\mathsf{Z}} f_i(z_i, y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i)$, where $\{z_i\}_{i=1}^n$ are the (unobserved) latent variables. In many applications of interest, the complete data likelihood belongs to the curved exponential family, *i.e.,*

$$f_i(z_i, y_i;\boldsymbol{\theta}) = h(z_i, y_i) \exp\left(\langle S_i(z_i, y_i) \,|\, \phi(\boldsymbol{\theta})\rangle - \psi(\boldsymbol{\theta})\right) \,, \quad (5.1.2)$$

where $\psi(\boldsymbol{\theta})$, $h(z_i, y_i)$ are scalar functions, $\phi(\boldsymbol{\theta}) \in \mathbb{R}^k$ is a vector function, and $S_i(z_i, y_i) \in \mathbb{R}^k$ is the complete data sufficient statistics. Latent variable models are widely used in machine learning and statistics; examples include mixture models for density estimation, clustering document, and topic modeling; see [McLachlan and Krishnan, 2007] and the references therein.

The basic "batch" EM (bEM) method iteratively computes a sequence of estimates $\{\boldsymbol{\theta}^k, k \in \mathbb{N}\}$ with an initial parameter $\boldsymbol{\theta}^0$. Each iteration of bEM is composed of two steps. In the E-step, a surrogate function is computed as $\boldsymbol{\theta} \mapsto Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{k-1}) = \sum_{i=1}^n Q_i(\boldsymbol{\theta}, \boldsymbol{\theta}^{k-1})$ where $Q_i(\boldsymbol{\theta}, \boldsymbol{\theta}') := -\int_{\mathsf{Z}} \log f_i(z_i, y_i; \theta) p_i(z_i|y_i;\boldsymbol{\theta}')\mu(\mathrm{d}z_i)$ such that $p_i(z_i|y_i;\boldsymbol{\theta}) := f_i(z_i, y_i;\boldsymbol{\theta})/g_i(y_i,\boldsymbol{\theta})$ is the conditional probability density of the latent variables z_i given the observations y_i. When $f_i(z_i, y_i;\boldsymbol{\theta})$ is a curved exponential family model, the E-step amounts to computing the conditional expectation of the complete data sufficient statistics,

$$\bar{\mathbf{s}}(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathbf{s}}_i(\boldsymbol{\theta}) \quad \text{where} \quad \bar{\mathbf{s}}_i(\boldsymbol{\theta}) = \int_{\mathsf{Z}} S_i(z_i, y_i) p_i(z_i|y_i;\boldsymbol{\theta})\mu(\mathrm{d}z_i) \,. \quad (5.1.3)$$

In the M-step, the surrogate function is minimized producing a new fit of the parameter $\boldsymbol{\theta}^k = \arg\max_{\boldsymbol{\theta}\in\Theta} Q(\boldsymbol{\theta}, \boldsymbol{\theta}^{k-1})$. The EM method has several appealing features – it is

monotone where the likelihood do not decrease at each iteration, invariant with respect to the parameterization, numerically stable when the optimization set is well defined, etc. The EM method has been the subject of considerable interest since its formalization in [Dempster et al., 1977].

With the sheer size of data sets today, the bEM method is not applicable as the E-step (5.1.3) involves a full pass over the dataset of n observations. Several approaches based on stochastic optimization have been proposed to address this problem. Neal and Hinton [1998] proposed (but not analyzed) an incremental version of EM, referred to as the iEM method. Cappé and Moulines [2009] developed the online EM (sEM) method which uses a stochastic approximation procedure to track the sufficient statistics defined in (5.1.3). Recently, Chen et al. [2018] proposed a variance reduced sEM (sEM-VR) method which is inspired by the SVRG algorithm popular in stochastic convex optimization [Johnson and Zhang, 2013]. The applications of the above stochastic EM methods are numerous, especially with the iEM and sEM methods; e.g., [Thiesson et al., 2001] for inference with missing data, [Ng and McLachlan, 2003] for mixture models and unsupervised clustering, [Hinton et al., 2006] for inference of deep belief networks, [Hofmann, 1999] for probabilistic latent semantic analysis, [Blei et al., 2017b, Wainwright et al., 2008] for variational inference of graphical models and [Ablin et al., 2018] for Independent Component Analysis.

This paper focuses on the theoretical aspect of stochastic EM methods by establishing novel *non-asymptotic* and *global* convergence rates for them. Our contributions are as follows.

- We offer two complementary views for the global convergence of EM methods – one focuses on the parameter space, and one on the sufficient statistics space. On one hand, the EM method can be studied as an *majorization-minimization* (MM) method in the parameter space. On the other hand, the EM method can be studied as a *scaled-gradient method* in the sufficient statistics space.
- Based on the two views described, we derive non-asymptotic convergence rate for stochastic EM methods. First, we show that the iEM method [Neal and Hinton, 1998] is a special instance of the MISO framework [Mairal, 2015a], and takes $\mathcal{O}(n/\epsilon)$ iterations to find an ϵ-stationary point to the ML estimation problem. Second, the sEM-VR method [Chen et al., 2018] is an instance of variance reduced stochastic scaled-gradient method, which takes $\mathcal{O}(n^{2/3}/\epsilon)$ iterations to find to an ϵ-stationary point.
- Lastly, we develop a Fast Incremental EM (fiEM) method based on the SAGA algorithm [Defazio et al., 2014, Reddi et al., 2016b] for stochastic optimization. We show that the new method is again a scaled-gradient method with the same iteration complexity as sEM-VR. This new method offers trade-off between storage cost and

computation complexity.

Importantly, our results capitalizes on the efficiency of stochastic EM methods applied on large datasets, and we support the above findings using numerical experiments.

Prior Work Since the empirical risk minimization problem (5.1.1) is typically *non-convex*, most prior work studying the convergence of EM methods considered either the *asymptotic* and/or *local* behaviors. For the classical study, the global convergence to a stationary point (either a local minimum or a saddle point) of the bEM method has been established by Wu et al. [1983] (by making the arguments developed in [Dempster et al., 1977] rigorous). The global convergence is a direct consequence of the EM method to be monotone. It is also known that in the neighborhood of a stationary point and under regularity conditions, the local rate of convergence of the bEM is linear and is given by the amount of *missing information* [McLachlan and Krishnan, 2007, Chapters 3 and 4].

The convergence of the iEM method was first tackled by Gunawardana and Byrne [2005] exploiting the interpretation of the method as an alternating minimization procedure under the information geometric framework developed in [Csiszár and Tusnády, 1984]. Although the EM algorithm is presented as an alternation between the E-step and M-step, it is also possible to take a variational perspective on EM to view both steps as maximization steps. Nevertheless, Gunawardana and Byrne [2005] assume that the latent variables take only a finite number of values and the order in which the observations are processed remains the same from one pass to the other.

More recently, the *local but non-asymptotic convergence* of EM methods has been studied in several works. These results typically require the initializations to be within a neighborhood of an isolated stationary point and the (negated) log-likelihood function to be strongly convex locally. Such conditions are either difficult to verify in general or have been derived only for specific models; see for example [Balakrishnan et al., 2017, Wang et al., 2015a, Xu et al., 2016a] and the references therein. The local convergence of sEM-VR method has been studied in [Chen et al., 2018, Theorem 1] but under a pathwise global stability condition.

5.2 Stochastic Optimization Techniques for EM Methods

We first describe the stochastic EM methods to be analyzed under a unified framework. The kth iteration of a generic stochastic EM method is composed of two sub-steps —

$$\text{sE-step}: \quad \hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} - \gamma_{k+1}\big(\hat{\mathbf{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\big), \tag{5.2.1}$$

which is a stochastic version of the E-step in (5.1.3). Note $\{\gamma_k\}_{k=1}^{\infty} \in [0,1]$ is a sequence of step sizes, $\boldsymbol{\mathcal{S}}^{(k+1)}$ is a proxy for $\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(k)})$, and $\bar{\mathbf{s}}$ is defined in (5.1.3). The M-step is given by

$$\text{M-step:}\quad \hat{\boldsymbol{\theta}}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\boldsymbol{s}}^{(k+1)}) := \operatorname*{arg\,min}_{\boldsymbol{\theta}\in\Theta} \{ \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \left\langle \hat{\boldsymbol{s}}^{(k+1)} \,|\, \phi(\boldsymbol{\theta}) \right\rangle \}, \tag{5.2.2}$$

which is controlled by the sufficient statistics determined by the sE-step. The stochastic EM methods differ in the way that $\boldsymbol{\mathcal{S}}^{(k+1)}$ is computed. Existing methods employ stochastic approximation or variance reduction without the need to fully compute $\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(k)})$. To simplify notations, we define

$$\bar{\mathbf{s}}_i^{(k)} := \bar{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}^{(k)}) = \int_{\mathsf{Z}} S(z_i, y_i) p_i(z_i|y_i; \hat{\boldsymbol{\theta}}^{(k)}) \mu(\mathrm{d}z_i) \quad \text{and} \quad \bar{\mathbf{s}}^{(\ell)} := \bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}^{(\ell)}) = \frac{1}{n}\sum_{i=1}^{n} \bar{\mathbf{s}}_i^{(\ell)}. \tag{5.2.3}$$

Note that if $\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}^{(k)}$ and $\gamma_{k+1} = 1$, eq. (5.2.1) reduces to the E-step in the classical bEM method. To describe the stochastic EM methods, let $i_k \in [\![1, n]\!]$ be a random index drawn at iteration k and $\tau_i^k = \max\{k' : i_{k'} = i,\ k' < k\}$ be the iteration index where $i \in [\![1, n]\!]$ is last drawn prior to iteration k, we have:

(iEM [Neal and Hinton, 1998])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \boldsymbol{\mathcal{S}}^{(k)} + \frac{1}{n}(\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\tau_{i_k}^k)})$	(5.2.4)
(sEM [Cappé and Moulines, 2009])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}_{i_k}^{(k)}$	(5.2.5)
(sEM-VR [Chen et al., 2018])	$\boldsymbol{\mathcal{S}}^{(k+1)} = \bar{\mathbf{s}}^{(\ell(k))} + (\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(\ell(k))})$	(5.2.6)

The stepsize is set to $\gamma_{k+1} = 1$ for the iEM method; $\gamma_{k+1} = \gamma$ is constant for the sEM-VR method. In the original version of the sEM method, the sequence of step γ_{k+1} is a diminishing step size. Moreover, for iEM we initialize with $\boldsymbol{\mathcal{S}}^{(0)} = \bar{\mathbf{s}}^{(0)}$; for sEM-VR, we set an epoch size of m and define $\ell(k) := m\lfloor k/m \rfloor$ as the first iteration number in the epoch that iteration k is in.

fiEM Our analysis framework can handle a new, yet natural application of a popular variance reduction technique to the EM method. The new method, called fiEM, is developed from the SAGA method [Defazio et al., 2014] in a similar vein as in sEM-VR.

For iteration $k \geq 0$, the fiEM method draws *two* indices *independently* and uniformly as $i_k, j_k \in [\![1, n]\!]$. In addition to τ_i^k which was defined *w.r.t.* i_k, we define $t_j^k = \{k' : j_{k'} = j, k' < k\}$ to be the iteration index where the sample $j \in [\![1, n]\!]$ is last drawn as j_k prior to iteration k. With the initialization $\overline{\boldsymbol{\mathcal{S}}}^{(0)} = \bar{\mathbf{s}}^{(0)}$, we use a slightly different update rule from SAGA inspired by [Reddi et al., 2016b], as described by the following recursive updates

$$\boldsymbol{\mathcal{S}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + (\bar{\mathbf{s}}_{i_k}^{(k)} - \bar{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}), \quad \overline{\boldsymbol{\mathcal{S}}}^{(k+1)} = \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}(\bar{\mathbf{s}}_{j_k}^{(k)} - \bar{\mathbf{s}}_{j_k}^{(t_{j_k}^k)}). \tag{5.2.7}$$

Algorithm 5.1 Stochastic EM methods.

Input: initializations $\hat{\boldsymbol{\theta}}^{(0)} \leftarrow 0$, $\hat{\mathbf{s}}^{(0)} \leftarrow \bar{\mathbf{s}}^{(0)}$, $K_{\mathsf{max}} \leftarrow$ max. iteration number.
Set the terminating iteration number, $K \in \{0, \ldots, K_{\mathsf{max}} - 1\}$, as a discrete r.v. with:

$$P(K = k) = \frac{\gamma_k}{\sum_{\ell=0}^{K_{\mathsf{max}}-1} \gamma_\ell}. \tag{5.2.8}$$

for $k = 0, 1, 2, \ldots, K$ **do**
  Draw index $i_k \in [\![1, n]\!]$ uniformly (and $j_k \in [\![1, n]\!]$ for fiEM).
  Compute the surrogate sufficient statistics $\boldsymbol{\mathcal{S}}^{(k+1)}$ using (5.2.5) or (5.2.4) or (5.2.6) or (5.2.7).
  Compute $\hat{\mathbf{s}}^{(k+1)}$ via the sE-step (5.2.1).
  Compute $\hat{\boldsymbol{\theta}}^{(k+1)}$ via the M-step (5.2.2).
end for
Return: $\hat{\boldsymbol{\theta}}^{(K)}$.

where we set a constant step size as $\gamma_{k+1} = \gamma$.

In the above, the update of $\boldsymbol{\mathcal{S}}^{(k+1)}$ corresponds to an *unbiased estimate* of $\bar{\mathbf{s}}^{(k)}$, while the update for $\overline{\boldsymbol{\mathcal{S}}}^{(k+1)}$ maintains the structure that $\overline{\boldsymbol{\mathcal{S}}}^{(k)} = n^{-1} \sum_{i=1}^{n} \bar{\mathbf{s}}_i^{(t_i^k)}$ for any $k \geq 0$. The two updates of (5.2.7) are based on two different and independent indices i_k, j_k that are randomly drawn from $[\![n]\!]$. This is used for our fast convergence analysis in Section 5.3.

We summarize the iEM, sEM-VR, sEM, fiEM methods in Algorithm 5.1. The random termination number (5.2.8) is inspired by [Ghadimi and Lan, 2013] which enables one to show non-asymptotic convergence to stationary point for non-convex optimization. Due to their stochastic nature, the per-iteration complexity for all the stochastic EM methods are independent of n, unlike the bEM method. They are thus applicable to large datasets with $n \gg 1$.

5.3 Global Convergence of Stochastic EM Methods

We establish non-asymptotic rates for the *global convergence* of the stochastic EM methods. We show that the iEM method is an instance of the incremental MM method; while sEM-VR, fiEM methods are instances of variance reduced *stochastic scaled gradient* methods. As we will see, the latter interpretation allows us to establish fast convergence rates of sEM-VR and fiEM methods.

First, we list a few assumptions which will enable the convergence analysis performed later in this section. Define:

$$\mathsf{S} := \left\{ \textstyle\sum_{i=1}^{n} \alpha_i \mathbf{s}_i \ : \ \mathbf{s}_i \in \operatorname{conv}\left\{ S(z, y_i) \ : \ z \in \mathsf{Z} \right\}, \ \alpha_i \in [-1, 1], \ i \in [\![1, n]\!] \right\}, \tag{5.3.1}$$

where $\operatorname{conv}\{A\}$ denotes the closed convex hull of the set A. From (5.3.1), we observe that

the iEM, sEM-VR, and fiEM methods generate $\hat{\boldsymbol{s}}^{(k)} \in \mathsf{S}$ for any $k \geq 0$. Consider:

H5.1 *The sets* Z, S *are compact. There exists constants* $C_{\mathsf{S}}, C_{\mathsf{Z}}$ *such that:*

$$C_{\mathsf{S}} := \max_{\mathbf{s},\mathbf{s}' \in \mathsf{S}} \|\mathbf{s} - \mathbf{s}'\| < \infty, \qquad C_{\mathsf{Z}} := \max_{i \in [\![1,n]\!]} \int_{\mathsf{Z}} |S_i(z, y_i)| \mu(\mathrm{d}z) < \infty. \tag{5.3.2}$$

H5.1 depends on the latent data model used and can be satisfied by several practical models (e.g., see Section 5.4). Denote by $\mathrm{J}_{\kappa}^{\boldsymbol{\theta}}(\boldsymbol{\theta}')$ the Jacobian of the function $\kappa : \boldsymbol{\theta} \mapsto \kappa(\boldsymbol{\theta})$ at $\boldsymbol{\theta}' \in \Theta$. Consider:

H5.2 *The function* ϕ *is smooth and bounded on* $\mathrm{int}(\Theta)$, i.e., *the interior of* Θ. *For all* $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathrm{int}(\Theta)^2$, $\| \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\boldsymbol{\theta}')\| \leq \mathrm{L}_{\phi} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$ *and* $\| \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\boldsymbol{\theta}')\| \leq C_{\phi}$.

H5.3 *The conditional distribution is smooth on* $\mathrm{int}(\Theta)$. *For any* $i \in [\![1, n]\!]$, $z \in \mathsf{Z}$, $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \mathrm{int}(\Theta)^2$, *we have* $\left|p_i(z|y_i; \boldsymbol{\theta}) - p_i(z|y_i; \boldsymbol{\theta}')\right| \leq \mathrm{L}_p \|\boldsymbol{\theta} - \boldsymbol{\theta}'\|$.

H5.4 *For any* $\boldsymbol{s} \in \mathsf{S}$, *the function* $\boldsymbol{\theta} \mapsto L(s, \boldsymbol{\theta}) := \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \langle \mathbf{s} \,|\, \phi(\boldsymbol{\theta}) \rangle$ *admits a unique global minimum* $\overline{\boldsymbol{\theta}}(\mathbf{s}) \in \mathrm{int}(\Theta)$. *In addition,* $\mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))$ *is full rank and* $\overline{\boldsymbol{\theta}}(\mathbf{s})$ *is* L_{θ}*-Lipschitz.*

Under H5.1, the assumptions H5.2 and H5.3 are standard for the curved exponential family distribution and the conditional probability distributions, respectively; H5.4 can be enforced by designing a strongly convex regularization function $\mathrm{R}(\boldsymbol{\theta})$ tailor made for Θ. We remark that for H5.3, it is possible to define the Lipschitz constant L_p independently for each data y_i to yield a refined characterization. We did not pursue such assumption to keep the notations simple.

Denote by $\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}, \boldsymbol{\theta})$ the Hessian (w.r.t to $\boldsymbol{\theta}$ for a given value of $\mathbf{s}$) of the function $\boldsymbol{\theta} \mapsto L(\mathbf{s}, \boldsymbol{\theta}) = \mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \langle \mathbf{s} \,|\, \phi(\boldsymbol{\theta}) \rangle$, and define

$$\mathrm{B}(\mathbf{s}) := \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s}, \overline{\boldsymbol{\theta}}(\mathbf{s})) \Big)^{-1} \mathrm{J}_{\phi}^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^{\top}. \tag{5.3.3}$$

H5.5 *It holds that* $\upsilon_{\max} := \sup_{\mathbf{s} \in \mathsf{S}} \| \mathrm{B}(\mathbf{s})\| < \infty$ *and* $0 < \upsilon_{\min} := \inf_{\mathbf{s} \in \mathsf{S}} \lambda_{\min}(\mathrm{B}(\mathbf{s}))$. *There exists a constant* L_B *such that for all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}^2$, *we have* $\| \mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\| \leq \mathrm{L}_B \|\mathbf{s} - \mathbf{s}'\|$.

Under H5.1, we have $\|\hat{\boldsymbol{s}}^{(k)}\| < \infty$ since S is compact. On the other hand, under H5.4, the EM methods generate $\hat{\boldsymbol{\theta}}^{(k)} \in \mathrm{int}(\Theta)$ for any $k \geq 0$. These assumptions ensure that the EM methods operate in a 'nice' set throughout the optimization process.

Detailed proofs for the theoretical results in this section are relegated to the appendix.

5.3.1 Incremental EM method

We show that the iEM method is a special case of the MISO method [Mairal, 2015a] utilizing the majorization minimization (MM) technique. The latter is a common technique for handling non-convex optimization. We begin by defining a surrogate function that majorizes $\mathcal{L}_i$:

$$Q_i(\boldsymbol{\theta};\boldsymbol{\theta}') := -\int_{\mathsf{Z}} \{\log f_i(z_i, y_i; \boldsymbol{\theta}) - \log p_i(z_i|y_i;\boldsymbol{\theta}')\}\, p_i(z_i|y_i;\boldsymbol{\theta}')\mu(\mathrm{d}z_i)\ . \tag{5.3.4}$$

The second term inside the bracket is a constant that does not depend on the first argument $\boldsymbol{\theta}$. Since $f_i(z_i, y_i; \boldsymbol{\theta}) = p_i(z_i|y_i;\boldsymbol{\theta})g_i(y_i;\boldsymbol{\theta})$, for all $\boldsymbol{\theta}' \in \Theta$, we get $Q_i(\boldsymbol{\theta}';\boldsymbol{\theta}') = -\log g_i(y_i;\boldsymbol{\theta}') = \mathcal{L}_i(\boldsymbol{\theta}')$. For all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, applying the Jensen inequality shows

$$Q_i(\boldsymbol{\theta},\boldsymbol{\theta}') - \mathcal{L}_i(\boldsymbol{\theta}) = \int \log \frac{p_i(z_i|y_i;\boldsymbol{\theta}')}{p_i(z_i|y_i;\boldsymbol{\theta})} p_i(z_i|y_i;\boldsymbol{\theta}')\mu(\mathrm{d}z_i) \geq 0 \tag{5.3.5}$$

which is the Kullback-Leibler divergence between the conditional distribution of the latent data $p(\cdot|y_i;\boldsymbol{\theta})$ and $p(\cdot|y_i;\boldsymbol{\theta}')$. Hence, for all $i \in [\![1, n]\!]$, $Q_i(\boldsymbol{\theta};\boldsymbol{\theta}')$ is a majorizing surrogate to $\mathcal{L}_i(\boldsymbol{\theta})$, *i.e.,* it satisfies for all $\boldsymbol{\theta}, \boldsymbol{\theta}' \in \Theta$, $Q_i(\boldsymbol{\theta};\boldsymbol{\theta}') \geq \mathcal{L}_i(\boldsymbol{\theta})$ with equality when $\boldsymbol{\theta} = \boldsymbol{\theta}'$. For the special case of curved exponential family distribution, the M-step of the iEM method is expressed as

$$\begin{aligned}\hat{\boldsymbol{\theta}}^{(k+1)} &\in \arg\min_{\boldsymbol{\theta}\in\Theta} \ \{\mathrm{R}(\boldsymbol{\theta}) + n^{-1}\textstyle\sum_{i=1}^n Q_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\tau_i^{(k+1)})})\} \\ &= \arg\min_{\boldsymbol{\theta}\in\Theta} \ \Big\{\mathrm{R}(\boldsymbol{\theta}) + \psi(\boldsymbol{\theta}) - \Big\langle n^{-1}\textstyle\sum_{i=1}^n \bar{\mathbf{s}}_i^{(\tau_i^{k+1})} \,|\, \phi(\boldsymbol{\theta})\Big\rangle)\Big\}.\end{aligned} \tag{5.3.6}$$

The iEM method can be interpreted through the MM technique — in the M-step, $\hat{\boldsymbol{\theta}}^{(k+1)}$ minimizes an upper bound of $\overline{\mathcal{L}}(\boldsymbol{\theta})$, while the sE-step updates the surrogate function in (5.3.6) which tightens the upper bound. Importantly, the error between the surrogate function and $\mathcal{L}_i$ is a smooth function:

Lemma 9 *Assume H5.1, H5.2, H5.3, H5.4. Let* $e_i(\boldsymbol{\theta};\boldsymbol{\theta}') := Q_i(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{L}_i(\boldsymbol{\theta})$. *For any* $(\boldsymbol{\theta}, \bar{\boldsymbol{\theta}}, \boldsymbol{\theta}') \in \Theta^3$, *we have* $\|\nabla e_i(\boldsymbol{\theta};\boldsymbol{\theta}') - \nabla e_i(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}')\| \leq \mathrm{L}_e \|\boldsymbol{\theta} - \bar{\boldsymbol{\theta}}\|$, *where* $\mathrm{L}_e := C_\phi C_{\mathsf{Z}} \mathrm{L}_p + C_{\mathsf{S}} \mathrm{L}_\phi$.

Proof The proof is postponed to Appendix 5.6

For *non-convex* optimization such as (5.1.1), it has been shown [Mairal, 2015a, Proposition 3.1] that the incremental MM method converges asymptotically to a stationary solution of a problem. We strengthen their result by establishing a non-asymptotic rate, which is new to the literature.

Theorem 5 *Consider the iEM algorithm,* i.e., *Algorithm 5.1 with (5.2.4). Assume H5.1, H5.2, H5.3, H5.4. For any $K_{\max} \geq 1$, it holds that*

$$\mathbb{E}[\|\nabla \overline{\mathcal{L}}(\boldsymbol{\theta}^{(K)})\|^2] \leq n \frac{2\,\mathrm{L}_e}{K_{\mathsf{max}}} \mathbb{E}[\overline{\mathcal{L}}(\boldsymbol{\theta}^{(0)}) - \overline{\mathcal{L}}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})], \tag{5.3.7}$$

where L_e is defined in Lemma 9 and K is a uniform random variable on $[\![0, K_{\max}-1]\!]$ [cf. (5.2.8)] independent of the $\{i_k\}_{k=0}^{K_{\max}}$.

Proof The proof is postponed to Appendix 5.7

We remark that under suitable assumptions, our analysis in Theorem 5 also extends to several non-exponential family distribution models.

5.3.2 Stochastic EM as Scaled Gradient Methods

We interpret the sEM-VR and fiEMmethods as *scaled gradient* methods on the sufficient statistics $\hat{\mathbf{s}}$, tackling a *non-convex* optimization problem. The benefit of doing so is that we are able to demonstrate a faster convergence rate for these methods through motivating them as *variance reduced* optimization methods. The latter is shown to be more effective when handling large datasets [Allen-Zhu and Hazan, 2016, Reddi et al., 2016a,b] than traditional stochastic/deterministic optimization methods. To set our stage, we consider the minimization problem:

$$\min_{\mathbf{s}\in\mathsf{S}} \; V(\mathbf{s}) := \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) = \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})), \tag{5.3.8}$$

where $\overline{\boldsymbol{\theta}}(\mathbf{s})$ is the unique map defined in the M-step (5.2.2). We first show that the stationary points of (5.3.8) has a one-to-one correspondence with the stationary points of (5.1.1):

Lemma 10 *For any $\mathbf{s} \in \mathsf{S}$, it holds that*

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{J}_{\overline{\boldsymbol{\theta}}}^{\mathbf{s}}(\mathbf{s})^{\top} \nabla_{\boldsymbol{\theta}} \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})). \tag{5.3.9}$$

Assume H5.4. If $\mathbf{s}^\star \in \{\mathbf{s} \in \mathsf{S} : \nabla_{\mathbf{s}} V(\mathbf{s}) = 0\}$, then $\overline{\boldsymbol{\theta}}(\mathbf{s}^\star) \in \left\{\boldsymbol{\theta} \in \Theta : \nabla_{\boldsymbol{\theta}} \overline{\mathcal{L}}(\boldsymbol{\theta}) = 0\right\}$. Conversely, if $\boldsymbol{\theta}^* \in \left\{\boldsymbol{\theta} \in \Theta : \nabla_{\boldsymbol{\theta}} \overline{\mathcal{L}}(\boldsymbol{\theta}) = 0\right\}$, then $\mathbf{s}^* = \bar{\mathbf{s}}(\boldsymbol{\theta}^*) \in \{\mathbf{s} \in \mathsf{S} : \nabla_{\mathbf{s}} V(\mathbf{s}) = 0\}$.

Proof The proof is postponed to Appendix 5.8

The next lemmas show that the update direction, $\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}$, in the sE-step (5.2.1) of sEM-VR and fiEM methods is a *scaled gradient* of $V(\mathbf{s})$. We first observe the following

conditional expectation:

$$\mathbb{E}\big[\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)} | \mathcal{F}_k\big] = \widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} = \hat{\mathbf{s}}^{(k)} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}}^{(k)})), \tag{5.3.10}$$

where $\mathcal{F}_k$ is the σ-algebra generated by $\{i_0, i_1, \ldots, i_k\}$ (or $\{i_0, j_0, \ldots, i_k, j_k\}$ for fiEM).

The difference vector $\mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))$ and the gradient vector $\nabla_{\mathbf{s}} V(\mathbf{s})$ are correlated, as we observe:

Lemma 11 *Assume H5.4,H5.5. For all* $\mathbf{s} \in \mathsf{S}$,

$$\upsilon_{\min}^{-1} \left\langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \right\rangle \geq \left\| \mathbf{s} - \overline{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \right\|^2 \geq \upsilon_{\max}^{-2} \|\nabla V(\mathbf{s})\|^2, \tag{5.3.11}$$

Proof The proof is postponed to Appendix 5.9

Combined with (5.3.10), the above lemma shows that the update direction in the sE-step (5.2.1) of sEM-VR and fiEM methods is a *stochastic scaled gradient* where $\hat{\mathbf{s}}^{(k)}$ is updated with a stochastic direction whose mean is correlated with $\nabla V(\mathbf{s})$.

Furthermore, the expectation step's operator and the objective function in (5.3.8) are smooth functions:

Lemma 12 *Assume H5.1, H5.3, H5.4, H5.5. For all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}$ *and* $i \in [\![1, n]\!]$, *we have*

$$\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \mathrm{L}_V \|\mathbf{s} - \mathbf{s}'\|, \tag{5.3.12}$$

where $\mathrm{L}_{\mathbf{s}} := C_{\mathsf{Z}} \mathrm{L}_p \mathrm{L}_\theta$ *and* $\mathrm{L}_V := \upsilon_{\max}(1 + \mathrm{L}_{\mathbf{s}}) + \mathrm{L}_B C_{\mathsf{S}}$.

Proof The proof is postponed to Appendix 5.10

The following theorem establishes the (fast) non-asymptotic convergence rates of sEM-VR and fiEM methods, which are similar to [Allen-Zhu and Hazan, 2016, Reddi et al., 2016a,b]:

Theorem 6 *Assume H5.1, H5.3, H5.4, H5.5. Denote* $\overline{L}_{\mathsf{v}} = \max\{\mathrm{L}_V, \mathrm{L}_{\mathbf{s}}\}$ *with the constants in Lemma 12.*

- *Consider the sEM-VR method,* i.e., *Algorithm 5.1 with (5.2.6). There exists a universal constant* $\mu \in (0, 1)$ *(independent of* n*) such that if we set the step size as* $\gamma = \frac{\mu \upsilon_{\min}}{\overline{L}_{\mathsf{v}} n^{2/3}}$ *and the epoch length as* $m = \frac{n}{2\mu^2 \upsilon_{\min}^2 + \mu}$, *then for any* $K_{\mathsf{max}} \geq 1$ *that is a multiple of* m, *it holds that*

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}} \frac{2\overline{L}_{\mathsf{v}}}{\mu K_{\mathsf{max}}} \frac{\upsilon_{\max}^2}{\upsilon_{\min}^2} \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(0)}) - V(\widehat{\boldsymbol{s}}^{(K_{\mathsf{max}})})]. \tag{5.3.13}$$

- *Consider the fiEM method,* i.e., *Algorithm 5.1 with (5.2.7). Set* $\gamma = \frac{\upsilon_{\min}}{\alpha \overline{L}_{\mathsf{v}} n^{2/3}}$ *such*

that $\alpha = \max\{6, 1 + 4\upsilon_{\min}\}$. *For any* $K_{\mathsf{max}} \geq 1$, *it holds that*

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}} \frac{\alpha^2 \overline{L}_{\mathsf{v}}}{K_{\mathsf{max}}} \frac{\upsilon_{\max}^2}{\upsilon_{\min}^2} \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(0)}) - V(\widehat{\boldsymbol{s}}^{(K_{\mathsf{max}})})]. \tag{5.3.14}$$

We recall that K *in the above is a uniform and independent r.v. chosen from* $[\![K_{\mathsf{max}} - 1]\!]$ *[cf. (5.2.8)].*

Proof The proof is postponed to Appendix 5.11

Comparing iEM, sEM-VR, and fiEM Note that by (5.3.9) in Lemma 10, if $\|\nabla_{\mathbf{s}} V(\hat{\mathbf{s}})\|^2 \leq \epsilon$, then $\|\nabla_{\boldsymbol{\theta}} \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}))\|^2 = \mathcal{O}(\epsilon)$, and vice versa, where the hidden constant is independent of n. In other words, the rates for iEM, sEM-VR, fiEM methods in Theorem 5 and 6 are comparable.

Importantly, the theorems show an intriguing comparison – to attain an ϵ-stationary point with $\|\nabla_{\boldsymbol{\theta}} \overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}))\|^2 \leq \epsilon$ or $\|\nabla_{\mathbf{s}} V(\hat{\mathbf{s}})\|^2 \leq \epsilon$, the iEM method requires $\mathcal{O}(n/\epsilon)$ iterations (in expectation) while the sEM-VR, fiEM methods require only $\mathcal{O}(n^{\frac{2}{3}}/\epsilon)$ iterations (in expectation). This comparison can be surprising since the iEM method is a monotone method as it guarantees decrease in the objective value; while the sEM-VR, fiEM methods are non-monotone. Nevertheless, it aligns with the recent analysis on stochastic variance reduction methods on non-convex problems. In the next section, we confirm the theory by observing a similar behavior numerically.

5.4 Application to Mixture and Topic Modeling

5.4.1 Gaussian Mixture Models

Our goal is to fit a GMM model to a set of n observations $\{y_i\}_{i=1}^n$ whose distribution is modeled as a Gaussian mixture of M components, each with a unit variance. Let $z_i \in [\![M]\!]$ be the latent labels, the complete log-likelihood is:

$$\log f_i(z_i, y_i; \boldsymbol{\theta}) = \textstyle\sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) \left[\log(\omega_m) - \mu_m^2/2\right] + \sum_{m=1}^M \mathbb{1}_{\{m\}}(z_i) \mu_m y_i + \text{constant} \,. \tag{5.4.1}$$

where $\boldsymbol{\theta} := (\boldsymbol{\omega}, \boldsymbol{\mu})$ with $\boldsymbol{\omega} = \{\omega_m\}_{m=1}^{M-1}$ are the mixing weights with the convention $\omega_M = 1 - \sum_{m=1}^{M-1} \omega_m$ and $\boldsymbol{\mu} = \{\mu_m\}_{m=1}^M$ are the means. We use the penalization $\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2} \sum_{m=1}^M \mu_m^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; M, \epsilon)$ where $\delta > 0$ and $\mathrm{Dir}(\cdot; M, \epsilon)$ is the M dimensional symmetric Dirichlet distribution with concentration parameter $\epsilon > 0$. The constraint set on $\boldsymbol{\theta}$ is given by

$$\Theta = \{\omega_m,\ m = 1, ..., M-1 : \omega_m \geq 0,\ \textstyle\sum_{m=1}^{M-1} \omega_m \leq 1\} \times \{\mu_m \in \mathbb{R},\ m = 1, ..., M\}. \tag{5.4.2}$$

In the following experiments of synthetic data, we generate samples from a GMM model with $M = 2$ components with two mixtures with means $\mu_1 = -\mu_2 = 0.5$, see Appendix 5.13.1 for details of the implementation and satisfaction of model assumptions for GMM inference.

Fixed sample size We use $n = 10^4$ synthetic samples and run the bEM method until convergence (to double precision) to obtain the ML estimate $\mu^\star$. We compare the bEM, sEM, iEM, sEM-VR and fiEM methods in terms of their precision measured by $|\mu - \mu^\star|^2$. We set the stepsize of the sEM as $\gamma_k = 3/(k+10)$, and the stepsizes of the sEM-VR and the fiEM to a constant stepsize proportional to $1/n^{2/3}$ and equal to $\gamma = 0.003$. The left plot of Figure 5.1 shows the convergence of the precision $|\mu - \mu^*|^2$ for the different methods against the epoch(s) elapsed (one epoch equals n iterations). We observe that the sEM-VR and fiEM methods outperform the other methods, supporting our analytical results.

Varying sample size We compare the number of *iterations* required to reach a precision of 10^{-3} as a function of the sample size from $n = 10^3$ to $n = 10^5$. We average over 5 independent runs for each method using the same stepsizes as in the finite sample size case above. The right plot of Figure 5.1 confirms our findings in Theorem 5 and 6. It requires $\mathcal{O}(n/\epsilon)$ (*resp.* $\mathcal{O}(n^{\frac{2}{3}}/\epsilon)$) iterations to find a ϵ-stationary point for the iEM (*resp.* sEM-VR and fiEM) method.

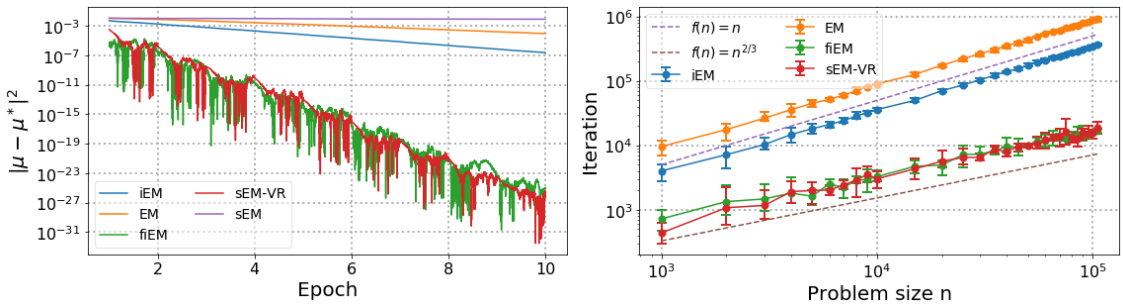

Figure 5.1 – Performance of stochastic EM methods for fitting a GMM. (Left) Precision ($|\mu^{(k)} - \mu^\star|^2$) as a function of the epoch elapsed. (Right) Number of iterations to reach a precision of 10^{-3}.

5.4.2 Probabilistic Latent Semantic Analysis

We are given a collection of documents $[\![D]\!]$ with terms from a vocabulary $[\![V]\!]$. The data is summarized by a list of tokens $\{y_i\}_{i=1}^n$ where each token is a pair of document and word $y_i = (y_i^{(\mathrm{d})}, y_i^{(\mathrm{w})})$ which indicates that $y_i^{(\mathrm{w})}$ appears in document $y_i^{(\mathrm{d})}$. The goal of pLSA is to classify the documents into K topics, which is modeled as a latent variable $z_i \in [\![K]\!]$ associated with each token [Hofmann, 1999].

Define $\boldsymbol{\theta} := (\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}, \boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})})$ as the parameter variable, where $\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})} = \{\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}_{d,k}\}_{[\![K-1]\!]\times[\![D]\!]}$ and $\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})} = \{\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}_{k,v}\}_{[\![K]\!]\times[\![V-1]\!]}$. The constraint set Θ is given as — for each $d \in [\![D]\!]$, $\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}_{d,\cdot} \in \Delta^K$ and for each $k \in [\![K]\!]$, we have $\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}_{\cdot,k} \in \Delta^V$, where Δ^K, Δ^V are the (reduced dimension) K, V-dimensional probability simplex; see (5.13.19) in the appendix for the precise definition.

Denote $\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}_{d,K} = 1 - \sum_{k=1}^{K-1} \boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}_{d,k}$ for each $d \in [\![D]\!]$, and $\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}_{k,V} = 1 - \sum_{\ell=1}^{V-1} \boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}_{k,\ell}$ for each $k \in [\![K]\!]$, the complete log likelihood for (y_i, z_i) is (up to an additive constant term):

$$\log f_i(z_i, y_i; \boldsymbol{\theta}) = \sum_{k=1}^{K}\sum_{d=1}^{D} \log(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}_{d,k}) \mathbb{1}_{\{k,d\}}(z_i, y_i^{(\mathrm{d})}) + \sum_{k=1}^{K}\sum_{v=1}^{V} \log(\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}_{k,v}) \mathbb{1}_{\{k,v\}}(z_i, y_i^{(\mathrm{w})}). \tag{5.4.3}$$

The penalization function is designed as

$$\mathrm{R}(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}, \boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}) = -\log \mathrm{Dir}(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})}; K, \alpha') - \log \mathrm{Dir}(\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})}; V, \beta'), \tag{5.4.4}$$

such that we ensure $\overline{\boldsymbol{\theta}}(\boldsymbol{s}) \in \mathrm{int}(\Theta)$. Lastly, the model assumptions and the implementation details are provided in Appendix 5.13.2.

Experiment We compare the EM methods on two FAO (UN Food and Agriculture Organization) datasets [Medelyan, 2009]. The first (*resp.* second) dataset consists of 10^3 (*resp.* 10.5×10^3) documents and a vocabulary of size 300. The number of topics is set to $K = 10$ and the stepsizes for the fiEM and sEM-VR are set to $\gamma = 1/n^{2/3}$ while the stepsize for the sEM is set to $\gamma_k = 1/(k + 10)$. Figure 5.1 shows the evidence lower bound (ELBO) as a function of the number of epochs for the datasets. Again, the result shows that fiEM and sEM-VR methods achieve faster convergence than the competing EM methods, affirming our theoretical findings.

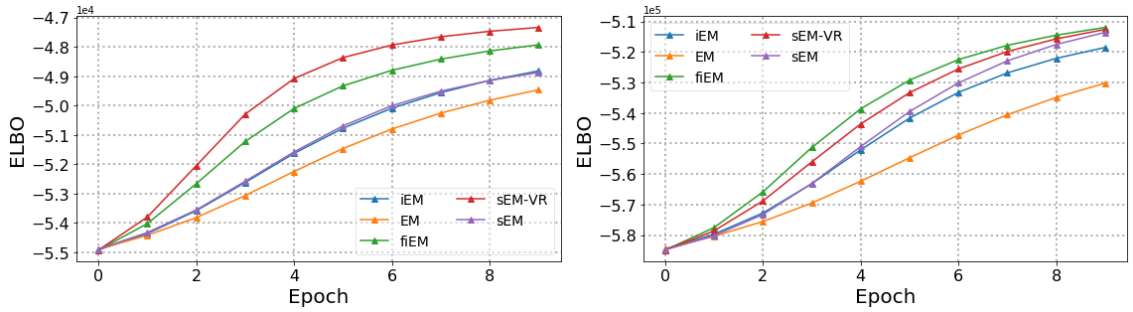

Figure 5.2 – ELBO of the stochastic EM methods on FAO datasets as a function of number of epochs elapsed. (Left) small dataset with 10^3 documents. (Right) large dataset with 10.5×10^3 documents.

5.5 Conclusion

This paper studies the global convergence for stochastic EM methods. Particularly, we focus on the inference of latent variable model with exponential family distribution and analyze the convergence of several stochastic EM methods. Our convergence results are *global* and *non-asymptotic*, and we offer two complimentary views on the existing stochastic EM methods — one interprets iEM method as an incremental MM method, and one interprets sEM-VR and fiEM methods as scaled gradient methods. The analysis shows that the sEM-VR and fiEM methods converge faster than the iEM method, and the result is confirmed via numerical experiments.

Appendices to Fast Incremental EM Methods

5.6 Proof of Lemma 9

Lemma *Assume H5.1, H5.2, H5.3, H5.4. Let* $e_i(\boldsymbol{\theta};\boldsymbol{\theta}') := Q_i(\boldsymbol{\theta};\boldsymbol{\theta}') - \mathcal{L}_i(\boldsymbol{\theta})$. *For any* $\boldsymbol{\theta},\bar{\boldsymbol{\theta}},\boldsymbol{\theta}' \in \Theta^3$, *we have* $\|\nabla e_i(\boldsymbol{\theta};\boldsymbol{\theta}') - \nabla e_i(\bar{\boldsymbol{\theta}};\boldsymbol{\theta}')\| \leq \mathrm{L}_e \|\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\|$, *where* $\mathrm{L}_e := C_\phi C_{\mathsf{Z}} \mathrm{L}_p + C_{\mathsf{S}} \mathrm{L}_\phi$.

Proof Observe the following identity

$$\begin{aligned}\nabla_{\boldsymbol{\theta}} \mathcal{L}_i(\boldsymbol{\theta})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} &= \nabla_{\boldsymbol{\theta}}\Big\{ -\log \int_{\mathsf{Z}} f_i(z_i, y_i; \boldsymbol{\theta}) \mu(\mathrm{d}z_i)\Big\}\Big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} \\ &\overset{(a)}{=} -\int_{\mathsf{Z}} \{\nabla_{\boldsymbol{\theta}} \log f_i(z_i, y_i; \boldsymbol{\theta})\}\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}} f_i(z_i|y_i;\hat{\boldsymbol{\theta}})\mu(\mathrm{d}z_i) \\ &= \nabla_{\boldsymbol{\theta}} Q_i(\boldsymbol{\theta};\hat{\boldsymbol{\theta}})\big|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\end{aligned} \tag{5.6.1}$$

where (a) is due to the Fisher's identity and (b) is due to the definition of Q_i in (5.3.4). It follows that

$$\nabla e_i(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}) = \nabla\{Q_i(\boldsymbol{\theta},\hat{\boldsymbol{\theta}}) - \mathcal{L}_i(\boldsymbol{\theta})\} = \mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})^\top (\bar{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}) - \bar{\mathbf{s}}_i(\boldsymbol{\theta})). \tag{5.6.2}$$

We observe that

$$\begin{aligned}\|\bar{\mathbf{s}}_i(\boldsymbol{\theta}) - \bar{\mathbf{s}}_i(\boldsymbol{\theta}')\| &= \Big\| \int_{\mathsf{Z}} S(z_i, y_i)\{p_i(z_i|y_i;\boldsymbol{\theta}) - p_i(z_i|y_i;\boldsymbol{\theta}')\}\mu(\mathrm{d}z_i)\Big\| \\ &\leq \mathrm{L}_p \|\boldsymbol{\theta}-\boldsymbol{\theta}'\| \int_{\mathsf{Z}} |S_i(z_i, y_i)|\mu(\mathrm{d}z_i) \leq C_{\mathsf{Z}} \mathrm{L}_p \|\boldsymbol{\theta}-\boldsymbol{\theta}'\|\end{aligned} \tag{5.6.3}$$

where the last inequality is due to the compactness of Z. Finally, we have

$$\begin{aligned}\|\nabla e_i(\boldsymbol{\theta};\hat{\boldsymbol{\theta}}) - \nabla e_i(\bar{\boldsymbol{\theta}};\hat{\boldsymbol{\theta}})\| &\leq \|\,\mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})\|\|\bar{\mathbf{s}}_i(\boldsymbol{\theta}) - \bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}})\| + \|\,\mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) - \mathrm{J}_\phi^{\boldsymbol{\theta}}(\bar{\boldsymbol{\theta}})\|\|\bar{\mathbf{s}}_i(\hat{\boldsymbol{\theta}}) - \bar{\mathbf{s}}_i(\bar{\boldsymbol{\theta}})\| \\ &\leq (C_\phi C_{\mathsf{Z}} \mathrm{L}_p + C_{\mathsf{S}} \mathrm{L}_\phi)\|\boldsymbol{\theta}-\bar{\boldsymbol{\theta}}\|\end{aligned} \tag{5.6.4}$$

where the last inequality is due to the compactness of S. ■

5.7 Proof of Theorem 5

Theorem *Consider the iEM algorithm,* i.e., *Algorithm 5.1 with* (5.2.4). *Assume H5.1, H5.2, H5.3, H5.4. For any* $K_{\max} \geq 1$, *it holds that*

$$\mathbb{E}[\|\nabla \overline{\mathcal{L}}(\boldsymbol{\theta}^{(K)})\|^2] \leq n \frac{2\,\mathrm{L}_e}{K_{\mathsf{max}}} \mathbb{E}[\overline{\mathcal{L}}(\boldsymbol{\theta}^{(0)}) - \overline{\mathcal{L}}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})],$$

where L_e *is defined in Lemma 9 and* K *is a uniform random variable on* $[\![0, K_{\max} - 1]\!]$ *[cf. (5.2.8)] independent of the* $\{i_k\}_{k=0}^{K_{\max}}$.

Proof We derive a *non-asymptotic* convergence rate for the iEM method. To begin our analysis, define

$$\overline{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) := \mathrm{R}(\boldsymbol{\theta}) + \frac{1}{n} \sum_{i=1}^{n} Q_i(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\tau_i^{k+1})}) \tag{5.7.1}$$

One has

$$\overline{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) = \overline{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n}(Q_{i_k}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(k)}) - Q_{i_k}(\boldsymbol{\theta}; \hat{\boldsymbol{\theta}}^{(\tau_{i_k}^k)})) \tag{5.7.2}$$

Observe that $\hat{\boldsymbol{\theta}}^{(k+1)} \in \arg\min_{\boldsymbol{\theta} \in \Theta} \overline{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta})$. We have

$$\begin{aligned} \overline{\mathcal{L}}^{(k+1)}(\hat{\boldsymbol{\theta}}^{(k+1)}) \leq \overline{\mathcal{L}}^{(k+1)}(\hat{\boldsymbol{\theta}}^{(k)}) &= \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) + \frac{1}{n}(Q_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(k)}) - Q_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(\tau_{i_k}^k)})) \\ &= \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) + \frac{1}{n}(\mathcal{L}_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}) - Q_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(\tau_{i_k}^k)})) \end{aligned} \tag{5.7.3}$$

where we have used the identity $\mathcal{L}_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}) = Q_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(k)})$. Arranging terms imply

$$e_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(\tau_{i_k}^k)}) = Q_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(\tau_{i_k}^k)}) - \mathcal{L}_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}) \leq n(\overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) - \overline{\mathcal{L}}^{(k+1)}(\hat{\boldsymbol{\theta}}^{(k+1)})) \tag{5.7.4}$$

For $k \in \mathbb{N}^*$, denote by $\mathcal{F}_k$ the σ-algebra generated by the random variables $i_0, \ldots, i_{k-1}$. Note that $\hat{\boldsymbol{\theta}}^{(k)}$ is $\mathcal{F}_k$-measurable. Because the random variable i_k is independent of $\mathcal{F}_{k-1}$ and is uniformly distributed over $\{1, \ldots, n\}$, the conditional expectation evaluates to

$$\mathbb{E}\left[e_{i_k}(\hat{\boldsymbol{\theta}}^{(k)}; \hat{\boldsymbol{\theta}}^{(\tau_{i_k}^k)}) \,\middle|\, \mathcal{F}_k \right] = \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) - \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)}) \tag{5.7.5}$$

where $\overline{\mathcal{L}}$ is the global objective function defined in (5.1.1). Note that the function $\overline{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \overline{\mathcal{L}}(\boldsymbol{\theta})$ is non-negative and L_e-smooth. It follows that for any $\boldsymbol{\theta}$, the inequality holds

$$0 \leq \overline{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \overline{\mathcal{L}}(\boldsymbol{\theta}) \leq \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) - \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)}) - \left\langle \nabla \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)}) \,|\, \boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(k)} \right\rangle + \frac{\mathrm{L}_e}{2} \|\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}^{(k)}\|^2, \tag{5.7.6}$$

where we have used the fact $\nabla \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) = \mathbf{0}$. Setting $\boldsymbol{\theta} = \hat{\boldsymbol{\theta}}^{(k)} + (\mathrm{L}_e)^{-1} \nabla \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)})$ in the

above yields

$$\frac{1}{2\,\mathrm{L}_e}\|\nabla\overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)})\|^2 \leq \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) - \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)}) \tag{5.7.7}$$

Therefore, taking the conditional expectation on both sides of (5.7.4) leads to

$$\frac{1}{2n\,\mathrm{L}_e}\|\nabla\overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)})\|^2 \leq \overline{\mathcal{L}}^{(k)}(\hat{\boldsymbol{\theta}}^{(k)}) - \mathbb{E}\big[\overline{\mathcal{L}}^{(k+1)}(\hat{\boldsymbol{\theta}}^{(k+1)})|\mathcal{F}_k\big] \tag{5.7.8}$$

Note that as we have set $\gamma_{k+1} = 1$ in the iEM method, the terminating iteration number K is chosen uniformly over $\{1, \ldots, K_{\mathsf{max}}\}$, therefore taking the total expectations gives

$$\begin{aligned}
\mathbb{E}[\|\nabla\overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(K)})\|^2] &= \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla\overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(k)})\|^2] \\
&\leq \frac{2n\,\mathrm{L}_e}{K_{\mathsf{max}}}\mathbb{E}\Big[\overline{\mathcal{L}}^{(0)}(\hat{\boldsymbol{\theta}}^{(0)}) - \overline{\mathcal{L}}^{(K_{\mathsf{max}})}(\hat{\boldsymbol{\theta}}^{(K_{\mathsf{max}}+1)})\Big] \\
&\leq \frac{2n\,\mathrm{L}_e}{K_{\mathsf{max}}}\mathbb{E}\Big[\overline{\mathcal{L}}^{(0)}(\hat{\boldsymbol{\theta}}^{(0)}) - \overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(K_{\mathsf{max}})})\Big]
\end{aligned} \tag{5.7.9}$$

Lastly, we note that $\overline{\mathcal{L}}(\hat{\boldsymbol{\theta}}^{(0)}) = \overline{\mathcal{L}}^{(0)}(\hat{\boldsymbol{\theta}}^{(0)})$. This leads to (5.3.7) and concludes our proof. ■

5.8 Proof of Lemma 10

Lemma *For any* $\mathbf{s} \in \mathsf{S}$*, it holds that*

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s})).$$

Assume H5.4. If $\mathbf{s}^\star \in \{\mathbf{s} \in \mathsf{S} : \nabla_{\mathbf{s}} V(\mathbf{s}) = 0\}$*, then* $\overline{\boldsymbol{\theta}}(\mathbf{s}^\star) \in \Big\{\boldsymbol{\theta} \in \Theta : \nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\boldsymbol{\theta}) = 0\Big\}$*. Conversely, if* $\boldsymbol{\theta}^* \in \Big\{\boldsymbol{\theta} \in \Theta : \nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\boldsymbol{\theta}) = 0\Big\}$*, then* $\mathbf{s}^* = \bar{\mathbf{s}}(\boldsymbol{\theta}^*) \in \{\mathbf{s} \in \mathsf{S} : \nabla_{\mathbf{s}} V(\mathbf{s}) = 0\}$.

Proof Using chain rule, we obtain $\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})^\top \nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))$ Obviously if $\nabla_{\mathbf{s}} V(\mathbf{s}^\star) = 0$, then $\nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\overline{\boldsymbol{\theta}}(\mathbf{s}^\star)) = 0$ because $\mathrm{J}_{\boldsymbol{\theta}}^{\mathbf{s}}(\mathbf{s})$ is invertible. Consider now the converse. By the Fisher identity, we get $\nabla_{\boldsymbol{\theta}}\mathcal{L}_i(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) - \mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})^\top \bar{\mathbf{s}}_i(\boldsymbol{\theta})$ which implies that $\nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\,\mathrm{R}(\boldsymbol{\theta}) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) - \mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})^\top \bar{\mathbf{s}}(\boldsymbol{\theta})$. Hence, if $\nabla_{\boldsymbol{\theta}}\overline{\mathcal{L}}(\boldsymbol{\theta}^*) = 0$, then $\nabla_{\boldsymbol{\theta}}\,\mathrm{R}(\boldsymbol{\theta}^*) + \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}^*) - \mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta}^n onb)^\top \bar{\mathbf{s}}^* = 0$ where we have set $\bar{\mathbf{s}}^* = \bar{\mathbf{s}}(\boldsymbol{\theta}^*)$. Under H5.4, the latter relation implies that $\boldsymbol{\theta}^* = \overline{\boldsymbol{\theta}}(\mathbf{s}^*)$. The proof follows. ■

5.9 Proof of Lemma 11

Lemma *Assume H5.4,H5.5. For all* $\mathbf{s} \in \mathsf{S}$*,*

$$\upsilon_{\min}^{-1}\Big\langle \nabla V(\mathbf{s}) \,|\, \mathbf{s} - \bar{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big\rangle \geq \big\|\mathbf{s} - \bar{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\big\|^2 \geq \upsilon_{\max}^{-2}\|\nabla V(\mathbf{s})\|^2,$$

Proof Using H5.4 and the fact that we can exchange integration with differentiation and the Fisher's identity, we obtain

$$\begin{aligned}\nabla_{\mathbf{s}} V(\mathbf{s}) &= \mathrm{J}^{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{s})^\top \Big(\nabla_{\boldsymbol{\theta}} \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathcal{L}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big) \\ &= \mathrm{J}^{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{s})^\top \Big(\nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big) \\ &= \mathrm{J}^{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{s})^\top \, \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \left(\mathbf{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \right) ,\end{aligned} \tag{5.9.1}$$

Consider the following vector map:

$$\mathbf{s} \to \nabla_{\boldsymbol{\theta}} L(\mathbf{s}, \boldsymbol{\theta})|_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\mathbf{s})} = \nabla_{\boldsymbol{\theta}} \psi(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \nabla_{\boldsymbol{\theta}} \mathrm{R}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top \mathbf{s} \, . \tag{5.9.2}$$

Taking the gradient of the above map *w.r.t.* $\mathbf{s}$ and using assumption H5.4, we show that:

$$\mathbf{0} = - \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\mathbf{s})) + \Big(\underbrace{\nabla^2_{\boldsymbol{\theta}} (\psi(\boldsymbol{\theta}) + \mathrm{R}(\boldsymbol{\theta}) - \langle \phi(\boldsymbol{\theta}) \,|\, \mathbf{s} \rangle)}_{= \mathrm{H}^{\boldsymbol{\theta}}_{L}(\mathbf{s}; \boldsymbol{\theta})} |_{\boldsymbol{\theta}=\overline{\boldsymbol{\theta}}(\mathbf{s})} \Big) \, \mathrm{J}^{\mathbf{s}}_{\boldsymbol{\theta}}(\mathbf{s}) \, . \tag{5.9.3}$$

The above yields

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{B}(\mathbf{s})(\mathbf{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))) \tag{5.9.4}$$

where we recall $\mathrm{B}(\mathbf{s}) = \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\mathbf{s})) \Big(\mathrm{H}^{\boldsymbol{\theta}}_{L}(\mathbf{s}; \overline{\boldsymbol{\theta}}(\mathbf{s})) \Big)^{-1} \mathrm{J}^{\boldsymbol{\theta}}_{\phi}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top$. The proof of (5.3.11) follows directly from the assumption H5.5. ■

5.10 Proof of Lemma 12

Lemma *Assume H5.1, H5.3, H5.4, H5.5. For all* $\mathbf{s}, \mathbf{s}' \in \mathsf{S}$ *and* $i \in [\![1, n]\!]$, *we have*

$$\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}} \|\mathbf{s} - \mathbf{s}'\|, \quad \|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \mathrm{L}_V \|\mathbf{s} - \mathbf{s}'\|,$$

where $\mathrm{L}_{\mathbf{s}} := C_{\mathsf{Z}} \, \mathrm{L}_p \, \mathrm{L}_\theta$ *and* $\mathrm{L}_V := \upsilon_{\max}(1 + \mathrm{L}_{\mathbf{s}}) + \mathrm{L}_B \, C_{\mathsf{S}}$.

Proof We prove the first inequality of the lemma in (5.3.12). Observe that

$$\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}')) = \int_{\mathsf{Z}} S(z_i, y_i) \{ p_i(z_i | y_i; \overline{\boldsymbol{\theta}}(\mathbf{s})) - p_i(z_i | y_i; \overline{\boldsymbol{\theta}}(\mathbf{s}')) \} \mu(\mathrm{d}z_i) \tag{5.10.1}$$

Taking norms on both sides and using H5.1, H5.3 yield

$$\|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_p \|\overline{\boldsymbol{\theta}}(\mathbf{s}) - \overline{\boldsymbol{\theta}}(\mathbf{s}')\| \int_{\mathsf{Z}} |S(z_i, y_i)| \mu(\mathrm{d}z_i) \leq C_{\mathsf{Z}} \, \mathrm{L}_p \|\overline{\boldsymbol{\theta}}(\mathbf{s}) - \overline{\boldsymbol{\theta}}(\mathbf{s}')\|, \tag{5.10.2}$$

where we have $\int_{\mathsf{Z}} |S(z_i, y_i)| \mu(\mathrm{d}z_i) \leq C_{\mathsf{Z}}$. Furthermore, under H5.4, as $\overline{\boldsymbol{\theta}}(\mathbf{s})$ is Lipschitz, there exists L_θ such that

$$\|\overline{\boldsymbol{\theta}}(\mathbf{s}) - \overline{\boldsymbol{\theta}}(\mathbf{s}')\| \leq \mathrm{L}_\theta \|\mathbf{s} - \mathbf{s}'\| \tag{5.10.3}$$

Substituting back into (5.10.2) concludes the proof with $\mathrm{L}_{\mathbf{s}} = C_{\mathsf{Z}}\,\mathrm{L}_p\,\mathrm{L}_\theta$.

To prove the second inequality in (5.3.12), we observe that:

$$\nabla_{\mathbf{s}} V(\mathbf{s}) = \mathrm{B}(\mathbf{s})\,(\mathbf{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))) \tag{5.10.4}$$

We observe the upper bound

$$\begin{aligned}
&\|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \\
&= \|\,\mathrm{B}(\mathbf{s})((\mathbf{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))) - (\mathbf{s}' - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}')))) + (\,\mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}'))(\mathbf{s}' - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})))\| \\
&\leq \|\,\mathrm{B}(\mathbf{s})\|\|\mathbf{s} - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - (\mathbf{s}' - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})))\| + \|\,\mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\|\|\mathbf{s}' - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\|
\end{aligned} \tag{5.10.5}$$

We observe that

$$\|\overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \frac{1}{n}\sum_{i=1}^{n} \|\overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s})) - \overline{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq \mathrm{L}_{\mathbf{s}}\,\|\mathbf{s} - \mathbf{s}'\|, \tag{5.10.6}$$

which is due to (5.3.12). Furthermore, as $\mathbf{s}' \in \mathsf{S}$, a compact set, we have $\|\mathbf{s}' - \overline{\boldsymbol{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}'))\| \leq C_{\mathsf{S}}$. Consequently, using H5.5 we have

$$\|\nabla V(\mathbf{s}) - \nabla V(\mathbf{s}')\| \leq \Big(\upsilon_{\max}(1 + \mathrm{L}_{\mathbf{s}}) + \mathrm{L}_B\,C_{\mathsf{S}}\Big)\|\mathbf{s} - \mathbf{s}'\|, \tag{5.10.7}$$

which proves our claim. ■

5.11 Proof of Theorem 6

Theorem *Assume H5.1, H5.3, H5.4, H5.5. Denote* $\overline{L}_{\mathsf{v}} = \max\{\mathrm{L}_V, \mathrm{L}_{\mathbf{s}}\}$ *with the constants in Lemma 12.*

[leftmargin=5.5mm]

- *Consider the sEM-VR method,* i.e., *Algorithm 5.1 with (5.2.6). There exists a universal constant* $\mu \in (0,1)$ *(independent of* n*) such that if we set the step size as* $\gamma = \frac{\mu \upsilon_{\min}}{\overline{L}_{\mathsf{v}} n^{2/3}}$ *and the epoch length as* $m = \frac{n}{2\mu^2 \upsilon_{\min}^2 + \mu}$*, then for any* $K_{\mathsf{max}} \geq 1$ *that is a multiple of* m*, it holds that*

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}}\,\frac{2\overline{L}_{\mathsf{v}}}{\mu K_{\mathsf{max}}}\,\frac{\upsilon_{\max}^2}{\upsilon_{\min}^2}\,\mathbb{E}[V(\widehat{\boldsymbol{s}}^{(0)}) - V(\widehat{\boldsymbol{s}}^{(K_{\mathsf{max}})})].$$

- *Consider the fiEM method,* i.e., *Algorithm 5.1 with (5.2.7). Set* $\gamma = \frac{\upsilon_{\min}}{\alpha \overline{L}_{\mathsf{v}} n^{2/3}}$ *such that* $\alpha = \max\{6, 1 + 4\upsilon_{\min}\}$*. For any* $K_{\mathsf{max}} \geq 1$*, it holds that*

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}}\,\frac{\alpha^2 \overline{L}_{\mathsf{v}}}{K_{\mathsf{max}}}\,\frac{\upsilon_{\max}^2}{\upsilon_{\min}^2}\,\mathbb{E}\big[V(\widehat{\boldsymbol{s}}^{(0)}) - V(\widehat{\boldsymbol{s}}^{(K_{\mathsf{max}})})\big].$$

We recall that K in the above is a uniform and independent r.v. chosen from $[\![K_{\mathsf{max}} - 1]\!]$ [cf. (5.2.8)].

To simplify notation, we shall denote $c_1 = \upsilon_{\min}^{-1}$ and $d_1 = \upsilon_{\max}$ in the below.

Proof for the sEM-VR method We first establish the following auxiliary lemma:

Lemma 13 *For any $k \geq 0$ and consider the update in (5.2.6), it holds that*

$$\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \leq 2\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\,\mathrm{L}_{\mathbf{s}}^2\,\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2], \tag{5.11.1}$$

where we recall that $\ell(k) := m\lfloor \frac{k}{m} \rfloor$ is the first iteration number in the epoch that iteration k is in.

Proof We observe that

$$\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \leq 2\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \tag{5.11.2}$$

For the latter term, we obtain its upper bound as

$$\begin{aligned}\mathbb{E}[\|\overline{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] &= \mathbb{E}\Big[\Big\|\frac{1}{n}\sum_{i=1}^{n}(\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{\ell(k)}) - (\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(\ell(k))})\Big\|^2\Big] \\ &\leq \mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(\ell(k))}\|^2] \leq \mathrm{L}_{\mathbf{s}}^2\,\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2]\end{aligned} \tag{5.11.3}$$

Substituting into (5.11.2) proves the lemma. ■

To proceed with our proof, we shall consider a constant step size $\gamma_k = \gamma$ and observe that

$$V(\widehat{\boldsymbol{s}}^{(k+1)}) \leq V(\widehat{\boldsymbol{s}}^{(k)}) - \gamma\left\langle \widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)} \,|\, \nabla V(\widehat{\boldsymbol{s}}^{(k)})\right\rangle + \frac{\gamma^2\,\mathrm{L}_V}{2}\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2 \tag{5.11.4}$$

Using (5.3.10) and taking expectations on both sides show that

$$\begin{aligned}&\mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k+1)})] \\ &\leq \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)})] - \gamma\mathbb{E}\Big[\left\langle \widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\widehat{\boldsymbol{s}}^{(k)})\right\rangle\Big] + \frac{\gamma^2\,\mathrm{L}_V}{2}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \\ &\overset{(a)}{\leq} \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)})] - \frac{\gamma}{c_1}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + \frac{\gamma^2\,\mathrm{L}_V}{2}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2]\end{aligned} \tag{5.11.5}$$

where (a) is due to Lemma 11. Furthermore, for $k+1 \leq \ell(k)+m$ (*i.e.,* $k+1$ is in the same epoch as k), we have

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2] = \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)} + \widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2] \\
&= \mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)}\|^2 + 2\left\langle \widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))} \mid \widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)} \right\rangle\Big] \\
&= \mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma^2\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2 - 2\gamma\left\langle \widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))} \mid \widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \right\rangle\Big] \\
&\leq \mathbb{E}\Big[(1+\gamma\beta)\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma^2\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2 + \frac{\gamma}{\beta}\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2\Big],
\end{aligned} \tag{5.11.6}$$

where the last inequality is due to the Young's inequality. Consider the following sequence

$$R_k := \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)}) + b_k\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2] \tag{5.11.7}$$

where $b_k := \overline{b}_{k \bmod m}$ is a periodic sequence where:

$$\overline{b}_i = \overline{b}_{i+1}(1+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2) + \gamma^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2, \quad i = 0, 1, \ldots, m-1 \;\text{ with }\; \overline{b}_m = 0. \tag{5.11.8}$$

Note that $\overline{b}_i$ is decreasing with i and this implies

$$\overline{b}_i \leq \overline{b}_0 = \gamma^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2\,\frac{(1+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2)^m - 1}{\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}, \; i = 1, 2, \ldots, m. \tag{5.11.9}$$

For $k+1 \leq \ell(k)+m$, we have the following inequality

$$\begin{aligned}
R_{k+1} &\leq \mathbb{E}\Big[V(\widehat{\boldsymbol{s}}^{(k)}) - \frac{\gamma}{c_1}\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 + \frac{\gamma^2\,\mathrm{L}_V}{2}\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2\Big] \\
&\quad + b_{k+1}\mathbb{E}\Big[(1+\gamma\beta)\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2 + \gamma^2\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2 + \frac{\gamma}{\beta}\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2\Big] \\
&\overset{(a)}{\leq} \mathbb{E}\Big[V(\widehat{\boldsymbol{s}}^{(k)}) - (\frac{\gamma}{c_1} - \frac{b_{k+1}\gamma}{\beta})\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 + b_{k+1}(1+\gamma\beta)\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2\Big] \\
&\quad + \Big(\gamma^2\,\mathrm{L}_V + 2b_{k+1}\gamma^2\Big)\mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 + \mathrm{L}_{\mathbf{s}}^2\,\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2\Big],
\end{aligned} \tag{5.11.10}$$

where (a) is due to Lemma 13. Rearranging terms gives

$$\begin{aligned}
R_{k+1} &\leq \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)})] - (\frac{\gamma}{c_1} - \frac{b_{k+1}\gamma}{\beta} - \gamma^2(\mathrm{L}_V + 2b_{k+1}))\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \\
&\quad + \Big(\underbrace{b_{k+1}(1+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2) + \gamma^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2}_{=b_k \quad \text{since } k+1 \leq \ell(k)+m}\Big)\mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(\ell(k))}\|^2\Big] \\
&= R_k - (\frac{\gamma}{c_1} - \frac{b_{k+1}\gamma}{\beta} - \gamma^2(\mathrm{L}_V + 2b_{k+1}))\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2]
\end{aligned} \tag{5.11.11}$$

This leads to, for any γ and β such that $(1 - c_1 b_{k+1}\beta^{-1} - c_1\gamma(\mathrm{L}_V + 2b_{k+1}) > 0$,

$$\frac{1}{d_1^2}\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(k)})\|^2] \leq \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \leq \frac{c_1(R_k - R_{k+1})}{\gamma\Big(1 - c_1 b_{k+1}\beta^{-1} - c_1\gamma(\mathrm{L}_V + 2b_{k+1})\Big)}. \quad (5.11.12)$$

By setting $\beta = \frac{c_1\overline{L}_{\mathsf{v}}}{n^{1/3}}$, $\gamma = \frac{\mu}{c_1\overline{L}_{\mathsf{v}}n^{2/3}}$, $m = \frac{nc_1^2}{2\mu^2 + \mu c_1^2}$, it can be shown that there exists $\mu \in (0,1)$, such that the following lower bound holds

$$\begin{aligned}
&1 - c_1\gamma\,\mathrm{L}_V - (\frac{c_1}{\beta} + 2c_1\gamma)b_{k+1} \geq 1 - \frac{\mu}{n^{\frac{2}{3}}} - \bar{b}_0(\frac{n^{\frac{1}{3}}}{\overline{L}_{\mathsf{v}}} + \frac{2\mu}{\overline{L}_{\mathsf{v}}n^{\frac{2}{3}}}) \\
&\geq 1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{\mathrm{L}_V\,\mu^2}{c_1^2 n^{\frac{4}{3}}}\frac{(1 + \gamma\beta + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2)^m - 1}{\gamma\beta + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}(\frac{n^{\frac{1}{3}}}{\overline{L}_{\mathsf{v}}} + \frac{2\mu}{\overline{L}_{\mathsf{v}}n^{\frac{2}{3}}}) \\
&\overset{(a)}{\geq} 1 - \frac{\mu}{n^{\frac{2}{3}}} - \frac{\mu}{c_1^2}(\mathrm{e} - 1)(1 + \frac{2\mu}{n}) \geq 1 - \mu - \mu(1 + 2\mu)\frac{\mathrm{e} - 1}{c_1^2} \overset{(b)}{\geq} \frac{1}{2}
\end{aligned} \quad (5.11.13)$$

where the simplification in (a) is due to

$$\frac{\mu}{n} \leq \gamma\beta + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2 \leq \frac{\mu}{n} + \frac{2\mu^2}{c_1^2 n^{\frac{4}{3}}} \leq \frac{\mu c_1^2 + 2\mu^2}{c_1^2}\frac{1}{n} \quad \text{and} \quad (1 + \gamma\beta + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2)^m \leq \mathrm{e} - 1. \quad (5.11.14)$$

and the required μ in (b) can be found by solving the quadratic equation [1]. This gives

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] = \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(k)})\|^2] \leq \frac{2d_1^2 c_1(R_0 - R_{K_{\mathsf{max}}})}{\gamma K_{\mathsf{max}}} \quad (5.11.15)$$

Note that $R_0 = \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(0)})]$ and if K_{max} is a multiple of m, then $R_{\mathsf{max}} = \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(K_{\mathsf{max}})})]$. Under the latter condition, we have

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] \leq n^{\frac{2}{3}}\,\frac{2d_1^2 c_1^2\overline{L}_{\mathsf{v}}}{\mu K_{\mathsf{max}}}\,\mathbb{E}[V(\widehat{\boldsymbol{s}}^{(0)}) - V(\widehat{\boldsymbol{s}}^{(K_{\mathsf{max}})})]. \quad (5.11.16)$$

This concludes our proof.

Proof for the fiEM method Our proof proceeds by observing the following auxiliary lemma:

Lemma 14 *For any $k \geq 0$ and consider the update in (5.2.7), it holds that*

$$\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \leq 2\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + \frac{2\,\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2] \quad (5.11.17)$$

Proof We observe that $\overline{\boldsymbol{\mathcal{S}}}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\overline{\boldsymbol{s}}_i^{(t_i^k)}$ and $\mathbb{E}[\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}] = \overline{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)}$. Moreover,

1. In fact, for small c_1, this gives $\mu = \Theta(c_1)$

we recall that $\overline{\boldsymbol{s}}_i^{(k)} = \overline{\boldsymbol{s}}_i(\boldsymbol{\theta}^{(k)}) = \overline{\boldsymbol{s}}_i(\overline{\boldsymbol{\theta}}(\widehat{\boldsymbol{s}}^{(k)}))$. Thus

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \overset{(a)}{=} \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} + (\overline{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)}) - (\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)})\|^2] \\
&\leq 2\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\mathbb{E}[\|(\overline{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{\mathcal{S}}}^{(k)}) - (\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)})\|^2] \\
&\overset{(b)}{\leq} 2\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + 2\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2],
\end{aligned} \tag{5.11.18}$$

where (a) uses the SAGA update in (5.2.7); (b) uses the variance inequality $\mathbb{E}[\|X - \mathbb{E}[X]\|^2] \leq \mathbb{E}[\|X\|^2]$. The last expectation can be further bounded by

$$\mathbb{E}[\|\overline{\boldsymbol{s}}_{i_k}^{(k)} - \overline{\boldsymbol{s}}_{i_k}^{(t_{i_k}^k)}\|^2] = \frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\overline{\boldsymbol{s}}_i^{(k)} - \overline{\boldsymbol{s}}_i^{(t_i^k)}\|^2] \overset{(a)}{\leq} \frac{\mathrm{L_s}}{n}\sum_{i=1}^n \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2], \tag{5.11.19}$$

where (a) is due to Lemma 11. Combining the two equations above yields the desired lemma. ■

Let $\gamma_{k+1} = \gamma$, *i.e.,* with a fixed step size. We observe the following

$$V(\widehat{\boldsymbol{s}}^{(k+1)}) \leq V(\widehat{\boldsymbol{s}}^{(k)}) - \gamma\left\langle \widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)} \,|\, \nabla V(\widehat{\boldsymbol{s}}^{(k)})\right\rangle + \frac{\gamma^2\,\mathrm{L}_V}{2}\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2 \tag{5.11.20}$$

Taking expectations on both sides yields

$$\begin{aligned}
&\mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k+1)})] \\
&\leq \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)})] - \gamma\mathbb{E}\Big[\left\langle \widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \nabla V(\widehat{\boldsymbol{s}}^{(k)})\right\rangle\Big] + \frac{\gamma^2\,\mathrm{L}_V}{2}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \\
&\overset{(a)}{\leq} \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)})] - \frac{\gamma}{c_1}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + \frac{\gamma^2\,\mathrm{L}_V}{2}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}\|^2] \\
&\overset{(b)}{\leq} \mathbb{E}[V(\widehat{\boldsymbol{s}}^{(k)})] - \Big(\frac{\gamma}{c_1} - \gamma^2\,\mathrm{L}_V\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + \frac{\gamma^2\,\mathrm{L}_V\,\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{i=1}^n \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2]
\end{aligned} \tag{5.11.21}$$

where (a) is due to Lemma 11 and (b) is due to Lemma 14. Next, we observe that

$$\frac{1}{n}\sum_{i=1}^n \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2] = \frac{1}{n}\sum_{i=1}^n \Big(\frac{1}{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)}\|^2] + \frac{n-1}{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2]\Big) \tag{5.11.22}$$

where the equality holds as i_k and j_k are drawn independently. For any $\beta > 0$, it holds

$$\begin{aligned}
&\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2] \\
&= \mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)}\|^2 + \|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2 + 2\left\langle \widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)} \,|\, \widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\right\rangle\Big] \\
&= \mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)}\|^2 + \|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2 - 2\gamma\left\langle \widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)} \,|\, \widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\right\rangle\Big] \\
&\leq \mathbb{E}\Big[\|\widehat{\boldsymbol{s}}^{(k+1)} - \widehat{\boldsymbol{s}}^{(k)}\|^2 + \|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2 + \frac{\gamma}{\beta}\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 + \gamma\beta\|\widehat{\boldsymbol{s}}^{(k)} - \widehat{\boldsymbol{s}}^{(t_i^k)}\|^2\Big]
\end{aligned} \tag{5.11.23}$$

where the last inequality is due to the Young's inequality. Subsequently, we have

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)}-\widehat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2]\\&\leq\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)}-\widehat{\boldsymbol{s}}^{(k)}\|^2]+\frac{n-1}{n^2}\sum_{i=1}^{n}\mathbb{E}\Big[(1+\gamma\beta)\|\widehat{\boldsymbol{s}}^{(k)}-\widehat{\boldsymbol{s}}^{(t_i^k)}\|^2+\frac{\gamma}{\beta}\|\widehat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2\Big]\end{aligned}\tag{5.11.24}$$

Observe that $\widehat{\boldsymbol{s}}^{(k+1)}-\widehat{\boldsymbol{s}}^{(k)}=-\gamma(\widehat{\boldsymbol{s}}^{(k)}-\boldsymbol{\mathcal{S}}^{(k+1)})$. Applying Lemma 14 yields

$$\begin{aligned}&\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)}-\widehat{\boldsymbol{s}}^{(t_i^{k+1})}\|^2]\\&\leq\Big(2\gamma^2+\frac{n-1}{n}\frac{\gamma}{\beta}\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]+\sum_{i=1}^{n}\Big(\frac{2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}{n}+\frac{(n-1)(1+\gamma\beta)}{n^2}\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)}-\widehat{\boldsymbol{s}}^{(t_i^k)}\|^2]\\&\leq\Big(2\gamma^2+\frac{\gamma}{\beta}\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]+\sum_{i=1}^{n}\frac{1-\frac{1}{n}+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)}-\widehat{\boldsymbol{s}}^{(t_i^k)}\|^2]\end{aligned}$$

Let us define

$$\Delta^{(k)}:=\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)}-\widehat{\boldsymbol{s}}^{(t_i^k)}\|^2]\tag{5.11.25}$$

From the above, we get

$$\Delta^{(k+1)}\leq\Big(1-\frac{1}{n}+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\Big)\Delta^{(k)}+\Big(2\gamma^2+\frac{\gamma}{\beta}\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)}-\overline{\boldsymbol{s}}^{(k)}\|^2]\tag{5.11.26}$$

Setting $\overline{L}_{\mathbf{v}}=\max\{\mathrm{L}_{\mathbf{s}},\mathrm{L}_V\}$, $\gamma=\frac{1}{\alpha c_1\overline{L}_{\mathbf{v}}n^{2/3}}$, $\beta=\frac{c_1\overline{L}_{\mathbf{v}}}{n^{1/3}}$, $(\alpha-1)c_1\geq 4$, $\alpha\geq 6$, it is easy to check that

$$1-\frac{1}{n}+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\geq 1-\frac{1}{n}\tag{5.11.27}$$

and

$$1-\frac{1}{n}+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\leq 1-\frac{1}{n}+\frac{1}{\alpha n}+\frac{2}{\alpha^2c_1^2n^{\frac{4}{3}}}\leq 1-\frac{\alpha c_1-c_1-2}{\alpha c_1 n}\leq 1-\frac{2}{\alpha c_1 n}\tag{5.11.28}$$

which shows that $1-\frac{1}{n}+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\in(0,1)$. Observe that as $\Delta^{(0)}=0$ and by telescoping, we have

$$\Delta^{(k+1)}\leq\Big(2\gamma^2+\frac{\gamma}{\beta}\Big)\sum_{\ell=0}^{k}\Big(1-\frac{1}{n}+\gamma\beta+2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\Big)^{k-\ell}\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(\ell)}-\overline{\boldsymbol{s}}^{(\ell)}\|^2]\tag{5.11.29}$$

Let $K_{\mathsf{max}} \in \mathbb{N}$. Summing $k = 0$ to $k = K_{\mathsf{max}} - 1$ gives

$$\begin{aligned}\sum_{k=0}^{K_{\mathsf{max}}-1} \Delta^{(k+1)} &\leq \Big(2\gamma^2 + \frac{\gamma}{\beta}\Big) \sum_{k=0}^{K_{\mathsf{max}}-1} \sum_{\ell=0}^{k} \Big(1 - \frac{1}{n} + \gamma\beta + 2\gamma^2\, \mathrm{L}_{\mathbf{s}}^2\Big)^{k-\ell} \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(\ell)} - \overline{\boldsymbol{s}}^{(\ell)}\|^2] \\ &= \Big(2\gamma^2 + \frac{\gamma}{\beta}\Big) \sum_{k=0}^{K_{\mathsf{max}}-1} \sum_{\ell=0}^{k} \Big(1 - \frac{1}{n} + \gamma\beta + 2\gamma^2\, \mathrm{L}_{\mathbf{s}}^2\Big)^{\ell} \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \\ &\leq \frac{2\gamma^2 + \frac{\gamma}{\beta}}{\frac{1}{n} - \gamma\beta - 2\gamma^2 L_{\mathbf{s}}^2} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2].\end{aligned} \tag{5.11.30}$$

Summing up the both sides of (5.11.21) from $k = 0$ to $k = K_{\mathsf{max}} - 1$ yields

$$\begin{aligned}&\mathbb{E}\big[V(\hat{\mathbf{s}}^{(K_{\mathsf{max}})}) - V(\hat{\mathbf{s}}^{(0)})\big] \\ &\leq \sum_{k=0}^{K_{\mathsf{max}}-1} \Big\{\Big(-\frac{\gamma}{c_1} + \gamma^2\, \mathrm{L}_V\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] + \gamma^2\, \mathrm{L}_V\, \mathrm{L}_{\mathbf{s}}^2\, \Delta^{(k)}\Big\} \\ &\leq \sum_{k=0}^{K_{\mathsf{max}}-1} \Big\{\Big(-\frac{\gamma}{c_1} + \gamma^2\, \mathrm{L}_V + \frac{(\gamma^2\, \mathrm{L}_V\, \mathrm{L}_{\mathbf{s}}^2)(2\gamma^2 + \frac{\gamma}{\beta})}{\frac{1}{n} - \gamma\beta - 2\gamma^2 L_{\mathbf{s}}^2}\Big)\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2]\Big\}\end{aligned} \tag{5.11.31}$$

Furthermore,

$$\begin{aligned}&\gamma^2\, \mathrm{L}_V + \frac{(\gamma^2\, \mathrm{L}_V\, \mathrm{L}_{\mathbf{s}}^2)(2\gamma^2 + \frac{\gamma}{\beta})}{\frac{1}{n} - \gamma\beta - 2\gamma^2 L_{\mathbf{s}}^2} \\ &\overset{(a)}{\leq} \frac{1}{\alpha^2 c_1^2 \overline{L}_{\mathsf{v}} n^{4/3}} + \frac{\overline{L}_{\mathsf{v}}(\alpha^2 c_1^2 n^{4/3})^{-1}(\frac{2}{\alpha^2 c_1^2 \overline{L}_{\mathsf{v}}^2 n^{4/3}} + \frac{1}{\alpha c_1^2 \overline{L}_{\mathsf{v}}^2 n^{1/3}})}{\frac{1}{n} - \frac{1}{\alpha n} - \frac{2}{\alpha^2 c_1^2 n^{4/3}}} \\ &= \frac{1}{\alpha^2 c_1^2 \overline{L}_{\mathsf{v}} n^{4/3}} + \frac{\overline{L}_{\mathsf{v}}(\frac{2}{\alpha^2 c_1^2 \overline{L}_{\mathsf{v}}^2 n^{4/3}} + \frac{1}{\alpha c_1^2 \overline{L}_{\mathsf{v}}^2 n^{1/3}})}{(\alpha c_1 n^{1/3})(\alpha - 1)c_1 - 2} \\ &\overset{(b)}{\leq} \frac{1}{\alpha^2 c_1^2 \overline{L}_{\mathsf{v}} n^{4/3}} + \frac{\frac{1}{\alpha c_1^2 \overline{L}_{\mathsf{v}} n^{1/3}}(\frac{2}{\alpha n} + 1)}{4(\alpha c_1 n^{1/3}) - 2} \overset{(c)}{\leq} \frac{1}{\alpha^2 c_1^2 \overline{L}_{\mathsf{v}} n^{4/3}} + \frac{2}{3\alpha^2 c_1^3 \overline{L}_{\mathsf{v}} n^{2/3}} \\ &\leq \frac{5/6}{\alpha c_1^2 \overline{L}_{\mathsf{v}} n^{2/3}}\end{aligned} \tag{5.11.32}$$

where (a) uses $\overline{L}_{\mathsf{v}} \geq \max\{\mathrm{L}_{\mathbf{s}}, \mathrm{L}_V\}$, (b) is due to $(\alpha - 1)c_1 \geq 4$ and (c) uses $\alpha c_1 n^{1/3} \geq 1$. Now, using the fact that $\frac{\gamma}{c_1} = \frac{1}{\alpha c_1^2 \overline{L}_{\mathsf{v}} n^{\frac{2}{3}}}$ and the lower bound $\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2 \geq d_2^{-1}\|\nabla V(\widehat{\boldsymbol{s}}^{(k)})\|^2$, we have

$$\begin{aligned}\mathbb{E}\big[V(\hat{\mathbf{s}}^{(K_{\mathsf{max}})}) - V(\hat{\mathbf{s}}^{(0)})\big] &\leq -\frac{1}{6\alpha c_1^2 \overline{L}_{\mathsf{v}} n^{\frac{2}{3}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \overline{\boldsymbol{s}}^{(k)}\|^2] \\ &\leq -\frac{1}{6\alpha d_1^2 c_1^2 \overline{L}_{\mathsf{v}} n^{\frac{2}{3}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(k)})\|^2]\end{aligned} \tag{5.11.33}$$

Recalling that K is an independent discrete r.v. drawn uniformly from $\{1, \dots, K_{\mathsf{max}}\}$ and noting that $\alpha \geq 6$, we have

$$\mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(K)})\|^2] = \frac{1}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\nabla V(\widehat{\boldsymbol{s}}^{(k)})\|^2] \leq n^{\frac{2}{3}} \frac{d_1^2 \overline{L}_{\mathsf{v}} (\alpha c_1)^2 (\mathbb{E}[V(\hat{\mathbf{s}}^{(0)}) - V(\hat{\mathbf{s}}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} \tag{5.11.34}$$

5.12 Local Linear Convergence of fiEM

In this section, we prove that the fiEM method converges locally at a linear rate to a stationary point, under a similar set of assumptions as in [Chen et al., 2018]. Note that some of the following assumptions can be difficult to verify, and our analysis here is merely a proof of concept.

Consider a stationary point $\boldsymbol{\theta}^\star$ to problem (5.1.1) and its corresponding sufficient statistics $\mathbf{s}^\star$, also a stationary point to (5.3.8). To simplify notations, we follow [Chen et al., 2018] and write the complete sufficient statistics as $F(\mathbf{s}') := \bar{\mathbf{s}}(\overline{\boldsymbol{\theta}}(\mathbf{s}'))$, and also the ith sufficient statistics as $f_i(\mathbf{s}') := \bar{\mathbf{s}}_i(\overline{\boldsymbol{\theta}}(\mathbf{s}'))$. We assume the following:

B 1 *The Hessian matrix $\nabla^2 \overline{\mathcal{L}}(\boldsymbol{\theta}^\star)$ is strictly positive definite such that $\boldsymbol{\theta}^\star$ is a strict local minimizer of problem (5.1.1).*

B 2 *For any $k \geq 1$, we have $\|\hat{\mathbf{s}}^k - \mathbf{s}^\star\| \leq \frac{\lambda}{\mathrm{L}_{\mathbf{s}}}$, where $\mathrm{L}_{\mathbf{s}}$ was defined in our Lemma 12 and $1 - \lambda$ is the maximum eigenvalue of the Jacobian matrix $\mathrm{J}_F^{\mathbf{s}}(\mathbf{s}^\star)$.*

The above assumptions correspond to assumptions (a), (c) in [Chen et al., 2018, Theorem 1], while we note that assumption (b) therein are shown in our Lemma 12.

We remark that B1 is strictly stronger than H5.4 used in our global convergence analysis. The latter makes assumption on the actual objective function $\overline{\mathcal{L}}(\boldsymbol{\theta}^\star)$ instead of the surrogate function $\boldsymbol{\theta} \to L(\mathbf{s}, \boldsymbol{\theta})$. Our proof goes as follows.

Proposition 12 *Under Assumption B1, B2 and the conditions such that our Lemma 12 holds. The fiEM method converges linearly such that*

$$\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \mathbf{s}^\star\|^2] \leq (1 - \delta)^{k+1} \|\widehat{\boldsymbol{s}}^{(0)} - \mathbf{s}^\star\|^2, \ \forall\ k \geq 0, \tag{5.12.1}$$

where $\delta = \Theta(1/n)$ with an appropriately chosen step size γ.

Proof (Sketch) For $k \in \mathbb{N}^*$, denote by $\mathcal{F}_k$ the σ-algebra generated by the random variables $i_0, j_0, \dots, i_k, j_k$. Consider

$$\begin{aligned} &\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \mathbf{s}^\star\|^2 | \mathcal{F}_k] = \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \gamma(\widehat{\boldsymbol{s}}^{(k)} - \boldsymbol{\mathcal{S}}^{(k+1)}) - \mathbf{s}^\star\|^2 | \mathcal{F}_k] \\ &= \mathbb{E}[\|(1-\gamma)\widehat{\boldsymbol{s}}^{(k)} + \gamma F(\widehat{\boldsymbol{s}}^{(k)}) - \mathbf{s}^\star + \gamma(\boldsymbol{\mathcal{S}}^{(k+1)} - F(\widehat{\boldsymbol{s}}^{(k)}))\|^2 | \mathcal{F}_k] \end{aligned} \tag{5.12.2}$$

Note that as $\mathbb{E}[\boldsymbol{\mathcal{S}}^{(k+1)} - F(\widehat{\boldsymbol{s}}^{(k)})|\mathcal{F}_k] = 0$, we have

$$\begin{aligned}&\mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k+1)} - \mathbf{s}^\star\|^2|\mathcal{F}_k] \\ &= \mathbb{E}[\|(1-\gamma)\widehat{\boldsymbol{s}}^{(k)} + \gamma F(\widehat{\boldsymbol{s}}^{(k)}) - \mathbf{s}^\star\|^2|\mathcal{F}_k] + \gamma^2\mathbb{E}[\|\boldsymbol{\mathcal{S}}^{(k+1)} - F(\widehat{\boldsymbol{s}}^{(k)})\|^2|\mathcal{F}_k]\end{aligned} \tag{5.12.3}$$

Repeating the analysis in (9) of [Chen et al., 2018], we arrive at the upper bound

$$\mathbb{E}[\|(1-\gamma)\widehat{\boldsymbol{s}}^{(k)} + \gamma F(\widehat{\boldsymbol{s}}^{(k)}) - \mathbf{s}^\star\|^2|\mathcal{F}_k] \leq (1-\gamma\lambda/2)\|\widehat{\boldsymbol{s}}^{(k)} - \mathbf{s}^\star\|^2 \tag{5.12.4}$$

On the other hand, applying [Defazio et al., 2014, Lemma 3] shows that

$$\begin{aligned}\mathbb{E}[\|\boldsymbol{\mathcal{S}}^{(k+1)} - F(\widehat{\boldsymbol{s}}^{(k)})\|^2|\mathcal{F}_k] &\leq 2\Big(\|f_{i_k}(\widehat{\boldsymbol{s}}^{(\tau_{i_k}^k)}) - f_{i_k}(\mathbf{s}^\star)\|^2 + \|f_{i_k}(\widehat{\boldsymbol{s}}^{(k)}) - f_{i_k}(\mathbf{s}^\star)\|^2\Big) \\ &\leq 2\,\mathrm{L}_{\mathbf{s}}^2\left(\|\widehat{\boldsymbol{s}}^{(\tau_{i_k}^k)} - \mathbf{s}^\star\|^2 + \|\widehat{\boldsymbol{s}}^{(k)} - \mathbf{s}^\star\|^2\right)\end{aligned} \tag{5.12.5}$$

Denote the total expectation as $h_k := \mathbb{E}[\|\widehat{\boldsymbol{s}}^{(k)} - \mathbf{s}^\star\|^2]$, and taking the total expectation on both sides yields

$$\mathbb{E}[\|\boldsymbol{\mathcal{S}}^{(k+1)} - F(\widehat{\boldsymbol{s}}^{(k)})\|^2] \leq 2\,\mathrm{L}_{\mathbf{s}}^2\left(h_k + \tfrac{1}{n}\textstyle\sum_{i=1}^n h_{\tau_i^k}\right) \tag{5.12.6}$$

Substituting the above into (5.12.3) yields

$$h_{k+1} \leq \Big(1 - \gamma\frac{\lambda}{2} + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\Big)h_k + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\left(\tfrac{1}{n}\textstyle\sum_{i=1}^n h_{\tau_i^k}\right) \tag{5.12.7}$$

Moreover, we observe the following recursion through evaluating the expectation

$$\frac{1}{n}\sum_{i=1}^n h_{\tau_i^k} = \frac{1}{n}h_{k-1} + \Big(1-\frac{1}{n}\Big)\frac{1}{n}\sum_{i=1}^n h_{\tau_i^{k-1}} \leq \frac{1}{n}\sum_{\ell=0}^{k-1}\Big(1-\frac{1}{n}\Big)^{k-\ell-1}h_\ell \tag{5.12.8}$$

Therefore, (5.12.7) simplifies to

$$h_{k+1} \leq \Big(1 - \gamma\frac{\lambda}{2} + 2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2\Big)h_k + \frac{2\gamma^2\,\mathrm{L}_{\mathbf{s}}^2}{n}\sum_{\ell=0}^{k-1}\Big(1-\frac{1}{n}\Big)^{k-\ell-1}h_\ell \tag{5.12.9}$$

To this end, we let $a = \frac{\lambda}{2}, b = 2\,\mathrm{L}_{\mathbf{s}}^2, c = 2\,\mathrm{L}_{\mathbf{s}}^2$ and consider the following inequality,

$$h_{k+1} \leq (1 - \gamma a + \gamma^2 b)h_k + \frac{\gamma^2 c}{n}\sum_{\ell=0}^{k-1}\Big(1-\frac{1}{n}\Big)^{k-\ell-1}h_\ell \tag{5.12.10}$$

We claim that for a sufficiently small step size γ, there exists $\delta \in (0,1]$ such that $h_k \leq (1-\delta)^k h_0$ for all k. The proof can be achieved using induction. The base case is straightforward since:

$$h_1 \leq (1 - \gamma a + \gamma^2 b)h_0 \tag{5.12.11}$$

For the induction case, we assume that $h_\tau \leq (1-\delta)^\tau h_0$ for $\tau = 1, 2...., k$. We observe that the induction hypothesis implies

$$\begin{aligned}
\frac{h_{k+1}}{h_0} &\leq (1-\gamma a+\gamma^2 b)(1-\delta)^k + \frac{\gamma^2 c}{n}\sum_{\ell=0}^{k-1}\Big(1-\frac{1}{n}\Big)^{k-\ell-1}(1-\delta)^\ell \\
&\leq (1-\gamma a+\gamma^2 b)(1-\delta)^k + \frac{\gamma^2 c}{n}(1-\delta)^{k-1}\frac{1}{1-\frac{1-1/n}{1-\delta}} \\
&= (1-\delta)^k\left\{(1-\gamma a+\gamma^2 b)+\gamma^2 c\frac{1}{1-n\delta}\right\} \\
&\overset{(a)}{\approx} (1-\delta)^k\left\{(1-\gamma a+\gamma^2 b)+\gamma^2 c(1+\delta n)\right\} \\
&\leq (1-\delta)^k\left\{(1-\gamma a+\gamma^2 b)+\gamma^2 c(1+n)\right\}
\end{aligned} \tag{5.12.12}$$

where the approximation holds if $n\delta \ll 1$. Lastly, if

$$\gamma \leq \frac{a}{2}(b+c(1+n))^{-1} \tag{5.12.13}$$

Then $h_{k+1} \leq (1-\delta)^{k+1}h_0$ with $\delta \leq \gamma a - \gamma^2(b+c(1+n)) = \mathcal{O}(1/n)$.

5.13 Practical Applications of Stochastic EM Methods

This section provides implementation details and verify the model assumptions for the application examples provided. Only in this section, for any $M \geq 2$, we denote

$$\Delta^M := \{\omega_m \in \mathbb{R},\ m = 1, ..., M-1 : \omega_m \geq 0,\ \textstyle\sum_{m=1}^{M-1}\omega_m \leq 1\} \subseteq \mathbb{R}^{M-1} \tag{5.13.1}$$

as the shorthand notation of the dimension reduced M-D probability simplex.

5.13.1 Gaussian mixture models

Model assumptions We first recognize that the constraint set for $\boldsymbol{\theta}$ is given by

$$\Theta = \Delta^M \times \mathbb{R}^M. \tag{5.13.2}$$

Using the partition of the sufficient statistics as $S(y_i, z_i) = (S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$, the partition $\phi(\boldsymbol{\theta}) = (\phi^{(1)}(\boldsymbol{\theta})^\top, \phi^{(2)}(\boldsymbol{\theta})^\top, \phi^{(3)}(\boldsymbol{\theta}))^\top \in \mathbb{R}^{M-1} \times \mathbb{R}^{M-1} \times \mathbb{R}$ and the fact that $\mathbb{1}_{\{M\}}(z_i) = 1 - \sum_{m=1}^{M-1}\mathbb{1}_{\{m\}}(z_i)$, the complete data log-likelihood can be expressed

as in (5.1.2) with

$$\begin{aligned} &s_{i,m}^{(1)} = \mathbb{1}_{\{m\}}(z_i), \quad \phi_m^{(1)}(\boldsymbol{\theta}) = \left\{\log(\omega_m) - \frac{\mu_m^2}{2}\right\} - \left\{\log(1 - \textstyle\sum_{j=1}^{M-1}\omega_j) - \frac{\mu_M^2}{2}\right\} , \\ &s_{i,m}^{(2)} = \mathbb{1}_{\{m\}}(z_i) y_i, \quad \phi_m^{(2)}(\boldsymbol{\theta}) = \mu_m , \\ &s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_M , \end{aligned} \tag{5.13.3}$$

and $\psi(\boldsymbol{\theta}) = -\left\{\log(1 - \sum_{m=1}^{M-1}\omega_m) - \frac{\mu_M^2}{2\sigma^2}\right\}$. We also define for each $m \in [\![1, M]\!]$, $j \in [\![1, 3]\!]$, $s_m^{(j)} = n^{-1}\sum_{i=1}^n s_{i,m}^{(j)}$. Consider the following conditional expected value:

$$\widetilde{\omega}_m(y_i; \boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}} | y = y_i] = \frac{\omega_m \exp(-\frac{1}{2}(y_i - \mu_i)^2)}{\sum_{j=1}^M \omega_j \exp(-\frac{1}{2}(y_i - \mu_j)^2)} , \tag{5.13.4}$$

where $m \in [\![1, M]\!]$, $i \in [\![1, n]\!]$ and $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\mu}) \in \Theta$. In particular, given $\boldsymbol{\theta} \in \Theta$, the E-step updates in (5.2.3) can be written as

$$\bar{\mathbf{s}}_i(\boldsymbol{\theta}) = (\underbrace{\widetilde{\omega}_1(y_i;\boldsymbol{\theta}), ..., \widetilde{\omega}_{M-1}(y_i;\boldsymbol{\theta})}_{:=\bar{\mathbf{s}}_i^{(1)}(\boldsymbol{\theta})^\top}, \underbrace{y_i\widetilde{\omega}_1(y_i;\boldsymbol{\theta}), ..., y_i\widetilde{\omega}_M(y_i;\boldsymbol{\theta})}_{:=\bar{\mathbf{s}}_i^{(2)}(\boldsymbol{\theta})^\top}, \underbrace{y_i}_{:=\bar{\mathbf{s}}_i^{(3)}(\boldsymbol{\theta})})^\top . \tag{5.13.5}$$

Recall that we have used the following regularizer:

$$\mathrm{R}(\boldsymbol{\theta}) = \tfrac{\delta}{2}\textstyle\sum_{m=1}^M \mu_m^2 - \epsilon\sum_{m=1}^M \log(\omega_m) - \epsilon\log\left(1 - \sum_{m=1}^{M-1}\omega_m\right) , \tag{5.13.6}$$

It can be shown that the regularized M-step in (5.2.5) evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1+\epsilon M)^{-1}(s_1^{(1)} + \epsilon, \ldots, s_{M-1}^{(1)} + \epsilon)^\top \\ ((s_1^{(1)} + \delta)^{-1} s_1^{(2)}, \ldots, (s_{M-1}^{(1)} + \delta)^{-1} s_{M-1}^{(2)})^\top \\ (1 - \sum_{m=1}^{M-1} s_m^{(1)} + \delta)^{-1}(s^{(3)} - \sum_{m=1}^{M-1} s_m^{(2)}) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_M(\boldsymbol{s}) \end{pmatrix} . \tag{5.13.7}$$

where we have defined for all $m \in [\![1, M]\!]$ and $j \in [\![1, 3]\!]$, $s_m^{(j)} = n^{-1}\sum_{i=1}^n s_{i,m}^{(j)}$.

To analyze the convergence of the EM methods, we verify H5.1 to H5.5 for the GMM example as follows.

To verify H5.1, we observe that the set Z is the compact interval $[\![M]\!]$, in addition, the sufficient statistics defined in (7.3.3) also leads to a bounded and closed S.

To verify H5.2, we observe that the Jacobian matrix $\mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta})$ can be computed as

$$\mathrm{J}_\phi^{\boldsymbol{\theta}}(\boldsymbol{\theta}) = \begin{pmatrix} \frac{1}{1-\sum_{m=1}^{M-1}\omega_m}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{\mathbf{1}}{\boldsymbol{\omega}}) & -\mathrm{Diag}(\boldsymbol{\mu}) & \mu_M\mathbf{1} \\ \mathbf{0} & \boldsymbol{I} & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & 1 \end{pmatrix} , \tag{5.13.8}$$

where we have denoted $\frac{\mathbf{1}}{\boldsymbol{\omega}}$ as the $(M-1)$-dimensional vector $(\frac{\mathbf{1}}{\omega_1}, \ldots, \frac{\mathbf{1}}{\omega_{M-1}})$. We observe that it is a bounded matrix and it is smooth *w.r.t.* $\boldsymbol{\theta}$.

We verify H5.3 next, *i.e.,* the Lipschitz continuity of $p(z_i|y_i;\boldsymbol{\theta})$, w.r.t to $\boldsymbol{\theta}$ noting that for all $i \in [\![n]\!]$ and $m \in [\![M]\!]$, $p(z_i = m|y_i;\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i=m\}}|y = y_i] = \widetilde{\omega}_m(y_i;\boldsymbol{\theta})$. Observe that $p(z_i = m|y_i;\boldsymbol{\theta})$ is given by the softmax function and the desired Lipschitz property follows.

Next, we observe that with the designed penalty, the function $\boldsymbol{\theta} \mapsto L(\mathbf{s},\boldsymbol{\theta})$ admits a unique global minima with $\overline{\boldsymbol{\theta}}(\mathbf{s}) \in \mathrm{int}(\Theta)$ for all $\mathbf{s} \in \mathsf{S}$. Second, since $\overline{\boldsymbol{\theta}}(\mathbf{s}) \in \mathrm{int}(\Theta)$, the Jacobian matrix defined in (5.13.8) must be full rank. Lastly, the L_θ-Lipschitzness of $\overline{\boldsymbol{\theta}}(\mathbf{s})$ can be deduced by inspecting (7.3.7). The above show that Assumption H5.4 is verified.

Finally, we calculate the quantity $\mathrm{B}(\mathbf{s})$ defined in (5.3.3). Observe that the Hessian $\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\theta})$ is:

$$\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\theta}) = \begin{pmatrix} \frac{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}{(1-\sum_{m=1}^{M-1}\omega_m)^2}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}{\boldsymbol{\omega}^2}) & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\boldsymbol{s}^{(1)}+\delta\mathbf{1}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \delta+1-\sum_{m=1}^{M-1} s_m^{(1)} \end{pmatrix} \tag{5.13.9}$$

We can rewrite $\mathrm{B}(\mathbf{s})$ as an outer product:

$$\mathrm{B}(\mathbf{s}) := \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top = \boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top \tag{5.13.10}$$

where

$$\boldsymbol{\mathcal{J}}(\boldsymbol{s}) := \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\boldsymbol{s}))\begin{pmatrix} \boldsymbol{H}_{11}^{-\frac{1}{2}} & \mathbf{0} & \mathbf{0} \\ \mathbf{0} & \mathrm{Diag}(\frac{\mathbf{1}}{\sqrt{\boldsymbol{s}^{(1)}+\delta\mathbf{1}}}) & \mathbf{0} \\ \mathbf{0} & \mathbf{0} & \frac{1}{\sqrt{\delta+1-\sum_{m=1}^{M-1} s_m^{(1)}}} \end{pmatrix} \tag{5.13.11}$$

and

$$\boldsymbol{H}_{11} := \frac{1+\epsilon-\sum_{m=1}^{M-1} s_m^{(1)}}{(1-\frac{\mathbf{1}^\top(\boldsymbol{s}^{(1)}+\epsilon\mathbf{1})}{1+\epsilon M})^2}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}(\frac{(1+\epsilon M)^2}{\boldsymbol{s}^{(1)}+\epsilon\mathbf{1}}). \tag{5.13.12}$$

Note that $\boldsymbol{\mathcal{J}}(\boldsymbol{s})$ is a bounded and full rank matrix which yields to the upper and lower bounds on eigenvealues in H5.5. From (5.13.11), we note that $\mathrm{B}(\mathbf{s}) = \boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top$ is Lipschitz continuous, *i.e.,* , there exists a constant L_B such that for all $\mathbf{s},\mathbf{s}' \in \mathsf{S}^2$, we have $\|\mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\| \leq \mathrm{L}_B\|\mathbf{s}-\mathbf{s}'\|$.

Algorithms updates In the sequel, for all $i \in [\![n]\!]$ and iteration k, the conditional expectation $\overline{\mathbf{s}}_i^{(k)}$ is defined by (7.3.5) and is equal to:

$$\overline{\mathbf{s}}_i^{(k)} = \begin{pmatrix} (\widetilde{\omega}_1(y_i;\boldsymbol{\theta}^{(k)}),\ldots,\widetilde{\omega}_{M-1}(y_i;\boldsymbol{\theta}^{(k)}))^\top \\ (y_i\widetilde{\omega}_1(y_i;\boldsymbol{\theta}^{(k)}),\ldots,y_i\widetilde{\omega}_{M-1}(y_i;\boldsymbol{\theta}^{(k)}))^\top \\ y_i \end{pmatrix} . \tag{5.13.13}$$

At iteration k, the several E-steps defined by (5.2.4) or (5.2.5) or (5.2.6) or (5.2.7) leads to the definition of the quantity $\hat{\mathbf{s}}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{\mathbf{s}}^{(0)} = n^{-1}\sum_{i=1}^n \overline{\mathbf{s}}_i^{(0)}$, those E-steps break down as follows:

Batch EM (EM): for all $i \in [\![1,n]\!]$, compute $\overline{\mathbf{s}}_i^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = n^{-1}\sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(k)} . \tag{5.13.14}$$

Online EM (sEM): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = (1-\gamma_k)\hat{\mathbf{s}}^{(k)} + \gamma_k \overline{\mathbf{s}}_{i_k}^{(k)} . \tag{5.13.15}$$

Incremental EM (iEM): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\tau_i^k)} = n^{-1}\sum\nolimits_{i=1}^n \overline{\mathbf{s}}_i^{(\tau_i^k)} . \tag{5.13.16}$$

Variance reduced stochastic EM (sEM-VR): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = (1-\gamma)\hat{\mathbf{s}}^{(k)} + \gamma\Big(\overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\ell(k))} + \overline{\mathbf{s}}^{(\ell(k))}\Big) \tag{5.13.17}$$

where $\overline{\mathbf{s}}_{i_k}^{(\ell(k))}$ and $\overline{\mathbf{s}}^{(\ell(k))}$ were computed at iteration $\ell(k)$, defined as the first iteration number in the epoch that iteration k is in.

Fast Incremental EM (fiEM): draw two different and independent indices (i_k, j_k) uniformly at random on $[\![n]\!]$, compute the quantities $\overline{\mathbf{s}}_{i_k}^{(k)}$ and $\overline{\mathbf{s}}_{j_k}^{(k)}$ and set

$$\begin{aligned} \hat{\mathbf{s}}^{(k+1)} &= (1-\gamma)\hat{\mathbf{s}}^{(k)} + \gamma\Big(\overline{\boldsymbol{\mathcal{S}}}^{(k)} + \overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(t_{i_k}^k)}\Big) \\ \overline{\boldsymbol{\mathcal{S}}}^{(k+1)} &= \overline{\boldsymbol{\mathcal{S}}}^{(k)} + n^{-1}\big(\overline{\mathbf{s}}_{j_k}^{(k)} - \overline{\mathbf{s}}_{j_k}^{(t_{j_k}^k)}\big) \end{aligned} \tag{5.13.18}$$

Finally, the k-th update reads $\boldsymbol{\theta}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$ where the function $\boldsymbol{s} \to \overline{\boldsymbol{\theta}}(\boldsymbol{s})$ is defined by (7.3.7).

5.13.2 Probabilistic Latent Semantic Analysis

Model assumptions The constraint set Θ is given by

$$\Theta = \left(\times_{d\in[\![D]\!]}\Delta^K\right) \times \left(\times_{k\in[\![K]\!]}\Delta^V\right). \tag{5.13.19}$$

For the sE-step (5.2.1) in the EM methods, we compute the expected complete data statistics as

$$\begin{aligned}
\bar{\mathbf{s}}^{(t|d)}_{i,d,k}(\boldsymbol{\theta}^{(\mathrm{t|d})}, \boldsymbol{\theta}^{(\mathrm{w|t})}) &= \mathbb{1}_{\{d\}}(y_i^{(\mathrm{d})})\Big(\textstyle\sum_{\ell=1}^K \boldsymbol{\theta}^{(\mathrm{t|d})}_{d,\ell}\boldsymbol{\theta}^{(\mathrm{w|t})}_{\ell,y_i^{(\mathrm{w})}}\Big)^{-1}\boldsymbol{\theta}^{(\mathrm{t|d})}_{d,k}\boldsymbol{\theta}^{(\mathrm{w|t})}_{k,y_i^{(\mathrm{w})}}, \\
\bar{\mathbf{s}}^{(w|t)}_{i,k,v}(\boldsymbol{\theta}^{(\mathrm{t|d})}, \boldsymbol{\theta}^{(\mathrm{w|t})}) &= \mathbb{1}_{\{v\}}(y_i^{(\mathrm{w})})\Big(\textstyle\sum_{\ell=1}^K \boldsymbol{\theta}^{(\mathrm{t|d})}_{y_i^{(\mathrm{d})},\ell}\boldsymbol{\theta}^{(\mathrm{w|t})}_{\ell,v}\Big)^{-1}\boldsymbol{\theta}^{(\mathrm{t|d})}_{y_i^{(\mathrm{d})},k}\boldsymbol{\theta}^{(\mathrm{w|t})}_{k,v},
\end{aligned} \tag{5.13.20}$$

for each $(i,k,d,v) \in [\![n]\!]\times[\![K]\!]\times[\![D]\!]\times[\![V]\!]$. Meanwhile, the regularized M-step (5.2.2) in the EM methods evaluates to:

$$\begin{pmatrix} \overline{\boldsymbol{\theta}}^{(\mathrm{t|d})}_{d,k}(\boldsymbol{s}) \\ \overline{\boldsymbol{\theta}}^{(\mathrm{w|t})}_{k,v}(\boldsymbol{s}) \end{pmatrix} = \begin{pmatrix} \Big(\sum_{i=1}^n\sum_{k'=1}^K \boldsymbol{s}^{(t|d)}_{i,d,k'} + \alpha' K\Big)^{-1}\big(\sum_{i=1}^n \boldsymbol{s}^{(t|d)}_{i,d,k} + \alpha'\big) \\ \Big(\sum_{i=1}^n\sum_{\ell=1}^V \boldsymbol{s}^{(w|t)}_{i,k,\ell} + \beta' V\Big)^{-1}\big(\sum_{i=1}^n \boldsymbol{s}^{(w|t)}_{i,k,v} + \beta'\big) \end{pmatrix}, \tag{5.13.21}$$

for each $(k,d,v) \in [\![K]\!]\times[\![D]\!]\times[\![V]\!]$.

Using the partition of the sufficient statistics as $S(y_i,z_i) = (S^{(t|d)}(y_i,z_i)^\top, S^{(w|t)}(y_i,z_i)^\top)^\top \in \mathbb{R}^{KD+KV}$, the partition $\phi(\boldsymbol{\theta}) = (\phi^{(t|d)}(\boldsymbol{\theta})^\top, \phi^{(w|t)}(\boldsymbol{\theta})^\top)^\top \in \mathbb{R}^{KD+KV}$, the complete log-likelihood (5.4.3) can be expressed in the standard form as (5.1.2) with

$$\begin{aligned}
\boldsymbol{s}^{(t|d)}_{i,d,k} &= \mathbb{1}_{\{k,d\}}(z_i, y_i^{(\mathrm{d})}), \quad \phi^{(t|d)}_{d,k}(\boldsymbol{\theta}) = \log(\boldsymbol{\theta}^{(\mathrm{t|d})}_{d,k}), \\
\boldsymbol{s}^{(w|t)}_{i,k,v} &= \mathbb{1}_{\{k,v\}}(z_i, y_i^{(\mathrm{w})}), \quad \phi^{(w|t)}_{k,v}(\boldsymbol{\theta}) = \log(\boldsymbol{\theta}^{(\mathrm{w|t})}_{k,v}),
\end{aligned} \tag{5.13.22}$$

Assumption H5.1 is verified with $\mathsf{Z} = [\![K]\!]$ and the sufficient statistics defined in (5.13.22) that leads to a compact S.

By using the vectorization of $\boldsymbol{\theta}$ as an $(K-1)D + (V-1)K$-dimensional vector, we can calculate the Jacobian as follows. In particular,

$$\mathrm{J}^{\phi^{(t|d)}_{d',k'}}_{\boldsymbol{\theta}^{(\mathrm{t|d})}_{d,k}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } d' \neq d, \\ \frac{1}{1-\sum_{\ell=1}^{K-1}\boldsymbol{\theta}^{(\mathrm{t|d})}_{d,\ell}} & \text{if } d' = d, k' \neq k \\ \frac{1}{\boldsymbol{\theta}^{(\mathrm{t|d})}_{d,k}} & \text{if } d' = d, k' = k. \end{cases}, \mathrm{J}^{\phi^{(w|t)}_{k',v'}}_{\boldsymbol{\theta}^{(\mathrm{w|t})}_{k,v}}(\boldsymbol{\theta}) = \begin{cases} 0 & \text{if } k' \neq k, \\ \frac{1}{1-\sum_{\ell=1}^{V-1}\boldsymbol{\theta}^{(\mathrm{w|t})}_{k,\ell}} & \text{if } k' = k, v' \neq v \\ \frac{1}{\boldsymbol{\theta}^{(\mathrm{w|t})}_{k,v}} & \text{if } k' = k, v' = v. \end{cases} \tag{5.13.23}$$

With the above definitions, it can be verified that the Jacobian matrix is full rank and smooth *w.r.t.* $\boldsymbol{\theta}$ for any $\boldsymbol{\theta} \in \mathrm{int}(\Theta)$. This confirms H5.2.

Next, we verify H5.3, *i.e.,* the Lipschitz continuity of $p(z_i|y_i;\boldsymbol{\theta})$, w.r.t to $\boldsymbol{\theta}$. Note that for all $(i,k,d) \in [\![n]\!] \times [\![K]\!] \times [\![D]\!]$, $p(z_i = k|y_i;\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})},\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})}) = \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{k,d\}}(z_i,y_i^{(\mathrm{d})})|y_i] = \bar{\mathbf{s}}_{i,k,d}^{(t|d)}(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})},\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})})$ as defined in (5.13.20). Observe that as we focus on $\boldsymbol{\theta} \in \mathrm{int}(\Theta)$, each of $\boldsymbol{\theta}_{d,\ell}^{(\mathrm{t}|\mathrm{d})}\boldsymbol{\theta}_{\ell,y_i^{(\mathrm{w})}}^{(\mathrm{w}|\mathrm{t})}$, $\boldsymbol{\theta}_{y_i^{(\mathrm{d})},\ell}^{(\mathrm{t}|\mathrm{d})}\boldsymbol{\theta}_{\ell,v}^{(\mathrm{w}|\mathrm{t})}$ is strictly positive and strictly less than one. The Lipschitz property follows from the expression (5.13.20).

The expression of the regularized complete log-likelihood, $\boldsymbol{\theta} \to L(s,\boldsymbol{\theta})$, is defined as:

$$L(s,\boldsymbol{\theta}) = -\sum_{k=1}^{K}\sum_{d=1}^{D} \boldsymbol{s}_{i,k,d}^{(t|d)} \log(\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})}) - \alpha' \log(\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})}) - \sum_{k=1}^{K}\sum_{v=1}^{V} \boldsymbol{s}_{i,k,v}^{(w|t)} \log(\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})}) - \beta' \log(\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})}) \ ,$$

This function admits a unique minimum in $\mathrm{int}(\Theta)$ from the strict concavity of the logarithm, as the regularizations are active with $\alpha', \beta' > 0$. By the same virtue of the verification of H5.2, we observe that H5.4 can be satisfied.

We first calculate the quantity $\mathrm{B}(\mathbf{s})$ defined in (5.3.3). Using the vectorization of $\boldsymbol{\theta}$ as a $(K-1)D + (V-1)K$-dimensional vector, we observe that the Hessian of the function $\boldsymbol{\theta} \mapsto L(\mathbf{s},\boldsymbol{\theta})$ *w.r.t.* to $\boldsymbol{\theta}$ has a block diagonal structure with $D+K$ blocks — the dth block which corresponds to $\boldsymbol{\theta}_{d,\cdot}^{(\mathrm{t}|\mathrm{d})}$ is given by

$$\big[\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\theta})\big]_d = \frac{\boldsymbol{s}_{K,d}^{(t|d)} + \alpha'}{(1-\sum_{k=1}^{K-1}\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^2}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}\Big(\frac{\boldsymbol{s}^{(t|d)} + \alpha'\mathbf{1}}{(\boldsymbol{\theta}^{(\mathrm{t}|\mathrm{d})})^2}\Big) \tag{5.13.24}$$

while the $(D+k)$th block which corresponds to $\boldsymbol{\theta}_{k,\cdot}^{(\mathrm{w}|\mathrm{t})}$ is given by

$$\big[\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\boldsymbol{\theta})\big]_{D+k} = \frac{\boldsymbol{s}_{k,V}^{(w|t)} + \beta'}{(1-\sum_{\ell=1}^{V-1}\boldsymbol{\theta}_{k,\ell}^{(\mathrm{w}|\mathrm{t})})^2}\mathbf{1}\mathbf{1}^\top + \mathrm{Diag}\Big(\frac{\boldsymbol{s}^{(w|t)} + \beta'\mathbf{1}}{(\boldsymbol{\theta}^{(\mathrm{w}|\mathrm{t})})^2}\Big) \tag{5.13.25}$$

Since each block in the above Hessian matrix is positive definite, the matrix

$$\mathrm{B}(\mathbf{s}) := \mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big(\mathrm{H}_L^{\boldsymbol{\theta}}(\mathbf{s},\overline{\boldsymbol{\theta}}(\mathbf{s}))\Big)^{-1}\mathrm{J}_\phi^{\boldsymbol{\theta}}(\overline{\boldsymbol{\theta}}(\mathbf{s}))^\top = \boldsymbol{\mathcal{J}}(\boldsymbol{s})\boldsymbol{\mathcal{J}}(\boldsymbol{s})^\top \tag{5.13.26}$$

is positive definite and bounded. Furthermore, there exists a constant L_B such that $\|\mathrm{B}(\mathbf{s}) - \mathrm{B}(\mathbf{s}')\| \leq \mathrm{L}_B\|\mathbf{s}-\mathbf{s}'\|$. Finally, this confirms H5.5.

Algorithms updates In the sequel, for all $(i,d,k,v) \in [\![n]\!] \times [\![D]\!] \times [\![K]\!] \times [\![V]\!]$ the conditional expectations $\bar{\mathbf{s}}_{i,k,d}^{(t|d)}((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^\delta, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^\delta)$ and $\bar{\mathbf{s}}_{i,k,v}^{(w|t)}((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^\delta, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^\delta)$ are defined by (5.13.20). For the pLSA example, after the initialization of the quantity $(\boldsymbol{s}_{k,d}^{(1)})^0 = n^{-1}\sum_{i=1}^n \bar{\mathbf{s}}_{i,k,d}^{(t|d)}((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^0, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^0)$ and $(\boldsymbol{s}_{k,v}^{(2)})^0 = n^{-1}\sum_{i=1}^n \bar{\mathbf{s}}_{i,k,v}^{(w|t)}((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^0, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^0)$, the several E-steps break down as follows:

Batch EM (EM): At iteration δ: update the statistics for all $(d,k,v) \in [\![D]\!] \times [\![K]\!] \times [\![V]\!]$

:

$$\big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta+1} = \sum_{i=1}^{n} \overline{\mathbf{s}}_{i,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) \quad \text{and} \quad \big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta+1} = \sum_{i=1}^{n} \overline{\mathbf{s}}_{i,k,v}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) \tag{5.13.27}$$

Online EM (sEM): At iteration δ, update the statistics for all $(d,k,v) \in [\![D]\!] \times [\![K]\!] \times [\![V]\!]$:

$$\begin{aligned} \big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta+1} &= (1-\gamma_\delta)\big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta} + \gamma_\delta \overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) \\ \big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta+1} &= (1-\gamma_\delta)\big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta} + \gamma_\delta \overline{\mathbf{s}}_{i_\delta,k,v}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) \end{aligned} \tag{5.13.28}$$

Incremental EM (iEM): At iteration δ, update the statistics for all $(d,k,v) \in [\![D]\!] \times [\![K]\!] \times [\![V]\!]$:

$$\begin{aligned} \big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta+1} &= \big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta} + \overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) - \overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\tau_{i_\delta}^{\delta}}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\tau_{i_\delta}^{\delta}}\big) \\ \big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta+1} &= \big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta} + \overline{\mathbf{s}}_{i_\delta,k,v}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) - \overline{\mathbf{s}}_{i_\delta,k,v}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\tau_{i_\delta}^{\delta}}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\tau_{i_\delta}^{\delta}}\big) \end{aligned} \tag{5.13.29}$$

Variance reduced stochastic EM (sEM-VR): At iteration δ, draw an index i_δ and update the statistics for all $(d,k,v) \in [\![D]\!] \times [\![K]\!] \times [\![V]\!]$:

$$\begin{aligned} &\big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta+1} = (1-\gamma)\big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta} \\ &+ \gamma\Big(\overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) - \overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{(\ell(k))}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{(\ell(k))}\big) + \overline{\mathbf{s}}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{(\ell(k))}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{(\ell(k))}\big)\Big) \\ &\big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta+1} = (1-\gamma)\big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta} \\ &+ \gamma\Big(\overline{\mathbf{s}}_{i_\delta,k,v}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{\delta}\big) - \overline{\mathbf{s}}_{i_\delta,k,v}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{(\ell(k))}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{(\ell(k))}\big) + \overline{\mathbf{s}}^{(w|t)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t}|\mathrm{d})})^{(\ell(k))}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w}|\mathrm{t})})^{(\ell(k))}\big)\Big) \end{aligned} \tag{5.13.30}$$

Fast Incremental EM (fiEM): At iteration δ, draw two indices (i_δ, j_δ) independently

and update the statistics for all $(d, k, v) \in [\![D]\!] \times [\![K]\!] \times [\![V]\!]$:

$$\big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta+1} = (1-\gamma)\big(\boldsymbol{s}_{k,d}^{(1)}\big)^{\delta} \tag{5.13.31}$$

$$+\gamma\Big(\overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{\delta}\big) - \overline{\mathbf{s}}_{i_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{(t_{i_\delta}^{\delta})}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{(t_{i_\delta}^{\delta})}\big) + \big(\overline{\boldsymbol{\mathcal{S}}}_{k,d}^{(1)}\big)^{\delta}\Big) \tag{5.13.32}$$

$$\big(\overline{\boldsymbol{\mathcal{S}}}_{k,d}^{(1)}\big)^{\delta+1} = \big(\overline{\boldsymbol{\mathcal{S}}}_{k,d}^{(1)}\big)^{\delta} + n^{-1}\Big(\overline{\mathbf{s}}_{j_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{\delta}\big) - \overline{\mathbf{s}}_{j_\delta,k,d}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{(t_{j_\delta}^{\delta})}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{(t_{j_\delta}^{\delta})}\big)\Big) \tag{5.13.33}$$

$$\big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta+1} = (1-\gamma)\big(\boldsymbol{s}_{k,v}^{(2)}\big)^{\delta} \tag{5.13.34}$$

$$+\gamma\Big(\overline{\mathbf{s}}_{i_\delta,k,v}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{\delta}\big) - \overline{\mathbf{s}}_{i_\delta,k,v}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{(t_{i_\delta}^{\delta})}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{(t_{i_\delta}^{\delta})}\big) + \big(\overline{\boldsymbol{\mathcal{S}}}_{k,v}^{(2)}\big)^{\delta}\Big) \tag{5.13.35}$$

$$\big(\overline{\boldsymbol{\mathcal{S}}}_{k,v}^{(2)}\big)^{\delta+1} = \big(\overline{\boldsymbol{\mathcal{S}}}_{k,v}^{(2)}\big)^{\delta} + \gamma n^{-1}\Big(\overline{\mathbf{s}}_{j_\delta,k,v}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{\delta}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{\delta}\big) - \overline{\mathbf{s}}_{j_\delta,k,v}^{(t|d)}\big((\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{(t_{j_\delta}^{\delta})}, (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{(t_{j_\delta}^{\delta})}\big)\Big) \tag{5.13.36}$$

Finally, at iteration δ, for $(k, d, v) \in [\![K]\!] \times [\![D]\!] \times [\![V]\!]$, the M-step in (5.2.2) evaluates to:

$$\begin{pmatrix} (\boldsymbol{\theta}_{d,k}^{(\mathrm{t|d})})^{\delta+1} \\ (\boldsymbol{\theta}_{k,v}^{(\mathrm{w|t})})^{\delta+1} \end{pmatrix} = \begin{pmatrix} \big(\sum_{k'=1}^{K} (\boldsymbol{s}_{k',d}^{(1)})^{\delta+1} + \alpha' K\big)^{-1}\big((\boldsymbol{s}_{k,d}^{(1)})^{\delta+1} + \alpha'\big) \\ \big(\sum_{\ell=1}^{V} (\boldsymbol{s}_{k,\ell}^{(2)})^{\delta+1} + \beta' V\big)^{-1}\big((\boldsymbol{s}_{k,v}^{(2)})^{\delta+1} + \beta'\big) \end{pmatrix}. \tag{5.13.37}$$

Chapter 6

Fast Stochastic Approximation of the EM

Abstract: *The ability to generate samples of the random effects from their conditional distributions is fundamental for inference in mixed effects models. Random walk Metropolis is widely used to perform such sampling, but this method is known to converge slowly for medium dimensional problems, or when the joint structure of the distributions to sample is spatially heterogeneous. The main contribution consists of an independent Metropolis-Hastings (MH) algorithm based on a multidimensional Gaussian proposal that takes into account the joint conditional distribution of the random effects and does not require any tuning. Indeed, this distribution is automatically obtained thanks to a Laplace approximation of the incomplete data model. Such approximation is shown to be equivalent to linearizing the structural model in the case of continuous data. Numerical experiments based on simulated and real data illustrate the performance of the proposed methods. For fitting nonlinear mixed effects models, the suggested MH algorithm is efficiently combined with a stochastic approximation version of the EM algorithm for maximum likelihood estimation of the global parameters. This chapter corresponds to the articles [Karimi and Lavielle, 2018] and [Karimi et al., 2020].*

Contents

6.1 Introduction

Mixed effects models are often adopted to take into account the inter-individual variability within a population (see [Lavielle, 2014] and the references therein). Consider a study with N individuals from a same population. The vector of observations y_i associated to each individual i is assumed to be a realisation of a random variable which depends on a vector of random individual parameters ψ_i. Then, inference on the individual parameter ψ_i amounts to estimate its conditional distribution given the observed data y_i.

When the model is a linear (mixed effects) Gaussian model, then this conditional distribution is a normal distribution that can explicitly be computed [Verbeke, 1997]. For more complex distributions and models, Monte Carlo methods must be used to approximate this conditional distribution. Most often, direct sampling from this conditional distribution is inefficient and it is necessary to resort to a Markov chain Monte Carlo (MCMC) method for obtaining random samples from this distribution. Yet, MCMC requires a tractable likelihood in order to compute the acceptance ratio. When this computation is impossible, a pseudo-marginal Metropolis Hastings (PMMH) has been developed in [Andrieu et al., 2009] and consists in replacing the posterior distribution evaluated in the MH acceptance rate by an unbiased approximation. An extension of the PMMH is the particle MCMC method, introduced in [Andrieu et al., 2010], where a Sequential Monte Carlo sampler [Doucet et al., 2000] is used to approximate the intractable likelihood at each iteration. For instance, this method is relevant when the model is SDE-based (see [Donnet and Samson, 2013]). In a fully Bayesian setting, approximation of the posterior of the global parameters can be used to approximate the posterior of the individual parameters using Integrated Nested Laplace Approximation (INLA) introduced in [Rue et al., 2009]. When the goal is to do approximate inference, this method has shown great performances mainly

because it approximates each marginal separately as univariate Gaussian distribution. In this paper, we focus on developing a method to perform exact inference and do not treat the case of approximate inference algorithms such as the Laplace EM or the First Order Conditional Estimation methods [Wang, 2007] that can introduce bias in the resulting parameters.

Note that generating random samples from $p_i(\psi_i|y_i;\boldsymbol{\theta})$ is useful for several tasks to avoid approximation of the model, such as linearisation or Laplace method. Such tasks include the estimation of the population parameters $\boldsymbol{\theta}$ of the model by a maximum likelihood approach, i.e. by maximizing the observed incomplete data likelihood $p(y_1,\dots y_N;\boldsymbol{\theta})$ using the Stochastic Approximation of the EM algorithm (SAEM) algorithm combined with a MCMC procedure [Kuhn and Lavielle, 2004]. Lastly, sampling from the conditional distributions $p_i(\psi_i|y_i;\boldsymbol{\theta})$ is also known to be useful for model building. Indeed, in [Lavielle and Ribba, 2016], the authors argue that methods for model assessment and model validation, whether graphical or based on statistical tests, must use samples of the conditional distribution $p_i(\psi_i|y_i;\boldsymbol{\theta})$ to avoid bias.

Designing a fast mixing sampler for these distributions is therefore of utmost importance to perform Maximum Likelihood Estimation (MLE) using the SAEM algorithm. The most common MCMC method for nonlinear mixed effects (NLME) models is the *random walk Metropolis* (RWM) algorithm [Lavielle, 2014, Robert and Casella, 2010, Roberts et al., 1997]. This method is implemented in software tools such as Monolix, NONMEM, the SAEMIX R package [Comets et al., 2017] and the nlmefitsa Matlab function. Despite its simplicity, it has been successfully used in many classical examples of pharmacometrics. Nevertheless, it can show its limitations when the dependency structure of the individual parameters is complex. Yet, maintaining an optimal acceptance rate (advocated in Roberts and Rosenthal [1997]) most often implies very small moves and therefore a very large number of iterations in medium and high dimensions since no information of the geometry of the target distribution is used.

The Metropolis-adjusted Langevin algorithm (MALA) uses evaluations of the gradient of the target density for proposing new states which are accepted or rejected using the Metropolis-Hastings algorithm [Roberts and Tweedie, 1996, Stramer and Tweedie, 1999]. Hamiltonian Monte Carlo (HMC) is another MCMC algorithm that exploits information about the geometry of the target distribution in order to efficiently explore the space by selecting transitions that can follow contours of high probability mass [Betancourt, 2017]. The No-U-Turn Sampler (NUTS) is an extension to HMC that allows an automatic and optimal selection of some of the settings required by the algorithm, [Brooks et al., 2011, Hoffman and Gelman, 2014]. Nevertheless, these methods may be difficult to use in practice, and are computationally involved, in particular when the structural model is a complex ODE based model. The algorithm we propose is an independent Metropolis-

Hastings (IMH) algorithm, but for which the proposal is a Gaussian approximation of the target distribution. For general data model (i.e. categorical, count or time-to-event data models or continuous data models), the Laplace approximation of the incomplete pdf $p_i(y_i;\boldsymbol{\theta})$ leads to a Gaussian approximation of the conditional distribution $p_i(\psi_i|y_i;\boldsymbol{\theta})$.

In the special case of continuous data, linearisation of the model leads, by definition, to a Gaussian linear model for which the conditional distribution of the individual parameter ψ_i given the data y_i is a multidimensional normal distribution that can be computed. Therefore, we design an independent sampler using this multivariate Gaussian distribution to sample from the target conditional distribution and embed this procedure in an exact inference algorithm, the SAEM, to speed the convergence of the vector of estimations of the global parameters $\hat{\boldsymbol{\theta}}$.

The paper is organised as follows. Mixed effects models for continuous and noncontinuous data are presented in Section 2. The standard MH for NLME models is described in Section 3. The proposed method, called the nlme-IMH, is introduced in Section 4 as well as the f-SAEM, a combination of this new method with the SAEM algorithm for estimating the population parameters of the model. Numerical examples illustrate, in Section 5, the practical performances of the proposed method, both on a continuous pharmacokinetics (PK) model and a time-to-event example. A Monte Carlo study confirms that this new SAEM algorithm shows a faster convergence to the maximum likelihood estimate.

6.2 Mixed Effect Models

6.2.1 Population approach and hierarchical models

In the sequel, we adopt a population approach, where we consider N individuals and n_i observations per individual i. The set of observed data is $y = (y_i, 1 \leq i \leq n)$ where $y_i = (y_{ij}, 1 \leq j \leq n_i)$ are the observations for individual i. For the sake of clarity, we assume that each observation y_{ij} takes its values in some subset of $\mathbb{R}$. The distribution of the n_i-vector of observations y_i depends on a vector of individual parameters ψ_i that takes its values in a subset of $\mathbb{R}^p$.

We assume that the pairs (y_i, ψ_i) are mutually independent and consider a parametric framework: the joint distribution of (y_i, ψ_i) is denoted by $f_i(y_i, \psi_i;\boldsymbol{\theta})$, where $\boldsymbol{\theta}$ is the vector of parameters of the model. A natural decomposition of this joint distribution reads

$$f_i(y_i, \psi_i;\boldsymbol{\theta}) = p_i(y_i|\psi_i;\boldsymbol{\theta})p_i(\psi_i;\boldsymbol{\theta}) \ , \tag{6.2.1}$$

where $p_i(y_i|\psi_i;\boldsymbol{\theta})$ is the conditional distribution of the observations given the individual parameters, and where $p_i(\psi_i;\boldsymbol{\theta})$ is the so-called population distribution used to describe

the distribution of the individual parameters within the population.

A particular case of this general framework consists in describing each individual parameter ψ_i as the sum of a typical value ψ_{pop} and a vector of individual random effects η_i:

$$\psi_i = \psi_{\text{pop}} + \eta_i \ . \tag{6.2.2}$$

In the sequel, we assume that the random effects are distributed according to a multivariate Gaussian distribution: $\eta_i \sim_{\text{i.i.d.}} \mathcal{N}(0, \Omega)$. Extensions of this general model are detailed in Appendix 6.7.1.

6.2.2 Continuous data models

A regression model is used to express the link between continuous observations and individual parameters:

$$y_{ij} = \mathsf{f}(t_{ij}, \psi_i) + \varepsilon_{ij} \ , \tag{6.2.3}$$

where y_{ij} is the j-th observation for individual i measured at index t_{ij}, ε_{ij} is the residual error. It is assumed that for any index t, $\psi \to \mathsf{f}(t, \psi)$ is twice differentiable in ψ.

We start by assuming that the residual errors are independent and normally distributed with zero-mean and a constant variance σ^2. Let $t_i = (t_{ij}, 1 \leq n_i)$ be the vector of observation indices for individual i. Then, the model for the observations reads:

$$y_i | \psi_i \sim \mathcal{N}(\mathsf{f}(\psi_i), \sigma^2 \mathtt{Id}_{n_i \times n_i}) \quad \text{where} \quad \mathsf{f}(\psi_i) = (\mathsf{f}(t_{i,1}, \psi_i), \ldots, \mathsf{f}(t_{i,n_i}, \psi_i)) \ . \tag{6.2.4}$$

If we assume that $\psi_i \sim_{\text{i.i.d.}} \mathcal{N}(\psi_{\text{pop}}, \Omega)$, then the parameters of the model are $\boldsymbol{\theta} = (\psi_{\text{pop}}, \Omega, \sigma^2)$.

Remark 6.1 *An extension of this model consists in assuming that the variance of the residual errors is not constant over time, i.e., $\varepsilon_{ij} \sim \mathcal{N}(0, g(t_{ij}, \psi_i)^2)$. Such extension includes proportional error models ($g = bf$) and combined error models ($g = a + bf$) [Lavielle, 2014] but the proposed method remains the same whatever the residual error model is.*

6.2.3 Noncontinuous data models

Noncontinuous data models include categorical data models [Agresti, 1990, Savic et al., 2011], time-to-event data models [Andersen, 2006, Mbogning et al., 2015], or count data models [Savic et al., 2011]. A categorical outcome y_{ij} takes its value in a set $\{1, \ldots, L\}$ of L categories. Then, the model is defined by the conditional probabilities

$(\mathbb{P}(y_{ij} = \ell|\psi_i), 1 \leq \ell \leq L)$, that depend on the vector of individual parameters ψ_i and may be a function of the time t_{ij}.

In a time-to-event data model, the observations are the times at which events occur. An event may be one-off (e.g., death, hardware failure) or repeated (e.g., epileptic seizures, mechanical incidents). To begin with, we consider a model for a one-off event. The survival function $S(t)$ gives the probability that the event happens after time t:

$$S(t) \triangleq \mathbb{P}(T > t) = \exp\left\{-\int_0^t h(u)\mathrm{d}u\right\} , \tag{6.2.5}$$

where h is called the hazard function. In a population approach, we consider a parametric and individual hazard function $h(\cdot, \psi_i)$. The random variable representing the time-to-event for individual i is typically written T_i and may possibly be right-censored. Then, the observation y_i for individual i is

$$y_i = \begin{cases} T_i & \text{if } T_i \leq \tau_c \\ \text{"}T_i > \tau_c\text{"} & \text{otherwise} , \end{cases} \tag{6.2.6}$$

where τ_c is the censoring time and "$T_i > \tau_c$" is the information that the event occurred after the censoring time.

For repeated event models, times when events occur for individual i are random times $(T_{ij}, 1 \leq j \leq n_i)$ for which conditional survival functions can be defined:

$$\mathbb{P}(T_{ij} > t|T_{i(j-1)} = t_{i(j-1)}) = \exp\left\{-\int_{t_{i(j-1)}}^t h(u, \psi_i)\mathrm{d}u\right\} . \tag{6.2.7}$$

Here, t_{ij} is the observed value of the random time T_{ij}. If the last event is right censored, then the last observation y_{i,n_i} for individual i is the information that the censoring time has been reached "$T_{i,n_i} > \tau_c$". The conditional pdf of $y_i = (y_{ij}, 1 \leq n_i)$ reads (see [Lavielle, 2014] for more details)

$$p_i(y_i|\psi_i) = \exp\left\{-\int_0^{\tau_c} h(u, \psi_i)\mathrm{d}u\right\} \prod_{j=1}^{n_i-1} h(t_{ij}, \psi_i) . \tag{6.2.8}$$

6.3 Sampling from Conditional Distributions

6.3.1 The conditional distribution of the individual parameters

Once the conditional distribution of the observations $p_i(y_i|\psi_i; \boldsymbol{\theta})$ and the marginal distribution of the individual parameters ψ_i are defined, the joint distribution $f_i(y_i, \psi_i; \boldsymbol{\theta})$ and the conditional distribution $p_i(\psi_i|y_i; \boldsymbol{\theta})$ are implicitly specified. This conditional distribution

$p_i(\psi_i|y_i;\boldsymbol{\theta})$ plays a crucial role for inference in NLME models.

One of the main task is to compute the maximum likelihood (ML) estimate of $\boldsymbol{\theta}$

$$\hat{\theta}_{\mathrm{ML}} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} \mathcal{L}(\boldsymbol{\theta}) \;, \tag{6.3.1}$$

where $\mathcal{L}(\boldsymbol{\theta}) = \log g(y;\boldsymbol{\theta})$. In NLME models, this optimization is solved by using a surrogate function defined as the conditional expectation of the complete data log-likelihood [McLachlan and Krishnan, 2007]. The SAEM is an iterative procedure for ML estimation that requires to generate one or several samples from this conditional distribution at each iteration of the algorithm. Once the ML estimate $\hat{\theta}_{\mathrm{ML}}$ has been computed, the observed Fisher information matrix noted $I(\hat{\theta}_{\mathrm{ML}}) = -\nabla^2_{\boldsymbol{\theta}}\mathcal{L}(\hat{\theta}_{\mathrm{ML}})$ can be derived thanks to the Louis formula [Louis, 1982] which expresses $I(\hat{\theta}_{\mathrm{ML}})$ in terms of the conditional expectation and covariance of the complete data log-likelihood. Such procedure also requires to sample from the conditional distributions $p_i(\psi_i|y_i;\hat{\theta}_{\mathrm{ML}})$ for all $i \in [\![1,n]\!]$.

Samples from the conditional distributions might also be used to define several statistical tests and diagnostic plots for models assessment. It is advocated in [Lavielle and Ribba, 2016] that such samples should be preferred to the modes of these distributions (also called *Empirical Bayes Estimate*(EBE), or *Maximum a Posteriori Estimate*), in order to provide unbiased tests and plots. For instance, a strong bias can be observed when the EBEs are used for testing the distribution of the parameters or the correlation between random effects.

In short, being able to sample individual parameters from their conditional distribution is essential in nonlinear mixed models. It is therefore necessary to design an efficient method to sample from this distribution.

6.3.2 The Metropolis-Hastings Algorithm

Metropolis-Hasting (MH) algorithm is a powerful MCMC procedure widely used for sampling from a complex distribution [Brooks et al., 2011]. To simplify the notations, we remove the dependency on $\boldsymbol{\theta}$. For a given individual $i \in [\![1,n]\!]$, the MH algorithm, to sample from the conditional distribution $p_i(\psi_i|y_i)$, is described as:

Algorithm 6.1 Metropolis-Hastings algorithm

Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution ξ_i .

Iteration k: given the current state of the chain $\psi_i^{(k-1)}$:

1. Sample a candidate ψ_i^c from a proposal distribution $q_i(\,\cdot\,|\psi_i^{(k-1)})$.
2. Compute the MH ratio:

$$\alpha(\psi_i^{(k-1)}, \psi_i^c) = \frac{p_i(\psi_i^c|y_i)}{p_i(\psi_i^{(k-1)}|y_i)} \frac{q_i(\psi_i^{(k-1)}|\psi_i^c)}{q_i(\psi_i^c|\psi_i^{(k-1)})} \,. \tag{6.3.2}$$

3. Set $\psi_i^{(k)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(k-1)})$ (otherwise, keep $\psi_i^{(k)} = \psi_i^{(k-1)}$).

Under weak conditions, $(\psi_i^{(k)}, k \geq 0)$ is an ergodic Markov chain whose distribution converges to the target $p_i(\psi_i|y_i)$ [Brooks et al., 2011].

Current implementations of the SAEM algorithm in Monolix [Chan et al., 2011], SAEMIX (R package) [Comets et al., 2017], nlmefitsa (Matlab) and NONMEM [Beal and Sheiner, 1980] mainly use the same combination of proposals. The first proposal is an independent MH algorithm which consists in sampling the candidate state directly from the prior distribution of the individual parameter ψ_i. The MH ratio then reduces to $p_i(y_i|\psi_i^c)/p_i(y_i|\psi_i^{(k)})$ for this proposal.

The other proposals are component-wise and block-wise random walk procedures [Metropolis et al., 1953] that updates different components of ψ_i using univariate and multivariate Gaussian proposal distributions. These proposals are centered at the current state with a diagonal variance-covariance matrix; the variance terms are adaptively adjusted at each iteration in order to reach some target acceptance rate [Atchadé et al., 2005, Lavielle, 2014]. Nevertheless, those proposals fail to take into account the nonlinear dependence structure of the individual parameters.

A way to alleviate these problems is to use a proposal distribution derived from a discretised Langevin diffusion whose drift term is the gradient of the logarithm of the target density leading to the Metropolis Adjusted Langevin Algorithm (MALA). The MALA proposal is a multivariate Gaussian with the following mean $\mu_{i,\text{MALA}}^{(k)}$ and covariance matrix Γ_{MALA}:

$$\mu_{i,\text{MALA}}^{(k)} = \psi_i^{(k)} + \gamma \nabla_{\psi_i} \log p_i(\psi_i^{(k)}|y_i) \quad \text{and} \quad \Gamma_{\text{MALA}} = 2\gamma \mathsf{I}_p \tag{6.3.3}$$

where γ is a positive stepsize and I_p is the identity matrix in $\mathbb{R}^{p\times p}$. These methods appear to behave well for complex models but still do not take into consideration the multidimensional structure of the individual parameters. Recent works include efforts in that direction, such as the Anisotropic MALA for which the covariance matrix of the proposal depends on the gradient of the target measure [Allassonniere and Kuhn, 2013], the Tamed Unadjusted Langevin Algorithm [Brosse et al., 2017] based on the coordinate-wise taming of superlinear drift coefficients and a multidimensional extension of the Adaptative

Metropolis algorithm [Haario et al., 2001] simultaneously estimating the covariance of the target measure and coercing the acceptance rate, see [Vihola, 2012].

The MALA algorithm is a special instance of the Hybrid Monte Carlo (HMC), introduced in [Neal et al., 2011]; see [Brooks et al., 2011] and the references therein, and consists in augmenting the state space with an auxiliary variable p, known as the velocity in Hamiltonian dynamics. This algorithm belongs to the class of data augmentation methods. Indeed, the potential energy is augmented with a kinetic energy, function of an added auxiliary variable. The MCMC procedure then consists in sampling from this augmented posterior distribution. All those methods aim at finding the proposal q that accelerates the convergence of the chain. Unfortunately they are computationally involved (even in small and medium dimension settings, the computation of the gradient or the Hessian can be overwhelming) and can be difficult to implement (stepsizes and numerical derivatives need to be tuned and implemented).

We see in the next section how to define a multivariate Gaussian proposal for both continuous and noncontinuous data models, that is easy to implement and that takes into account the multidimensional structure of the individual parameters in order to accelerate the MCMC procedure.

6.4 The nlme-IMH and the f-SAEM

In this section, we assume that the individual parameters $(\psi_1, \ldots, \psi_n)$ are independent and normally distributed with mean $(m_1, \ldots, m_N)$ and covariance Ω. The MAP estimate, for individual i, is the value of ψ_i that maximizes the conditional distribution $p_i(\psi_i|y_i, \boldsymbol{\theta})$:

$$\hat{\psi}_i = \arg\max_{\psi_i \in \mathbb{R}^p} p_i(\psi_i|y_i) = \arg\max_{\psi_i \in \mathbb{R}^p} p_i(y_i|\psi_i)p_i(\psi_i) \ . \tag{6.4.1}$$

6.4.1 Proposal based on Laplace approximation

For both continuous and noncontinuous data models, the goal is to find a simple proposal, a multivariate Gaussian distribution in our case, that approximates the target distribution $p_i(\psi_i|y_i)$. For general MCMC samplers, it is shown in [Roberts and Rosenthal, 2011] that the mixing rate in total variation depends on the expectation of the acceptance ratio under the proposal distribution which is also directly related to the ratio of the proposal to the target in the special case of independent samplers (see [Mengersen and Tweedie, 1996, Roberts and Rosenthal, 2011]). This observation naturally suggests to find a proposal which approximates the target. de Freitas et al. [2001] advocates the use a multivariate Gaussian distribution whose parameters are obtained by minimizing the Kullback-Leibler divergence between a multivariate Gaussian variational candidate

distribution and the target distribution. In [Andrieu and Thoms, 2008] and the references therein, an adaptative Metropolis algorithm is studied and reconciled to a KL divergence minimisation problem where the resulting multivariate Gaussian distribution can be used as a proposal in a IMH algorithm. Authors note that although this proposal might be a sensible choice when it approximates well the target, it can fail when the parametric form of the proposal is not sufficiently rich. Thus, other parametric forms can be considered and it is suggested in [Andrieu et al., 2006] to consider mixtures, finite or infinite, of distributions belonging to the exponential family.

In general, this optimization step is difficult and computationally expensive since it requires to approximate (using Monte Carlo integration for instance) the integral of the log-likelihood with respect to the variational candidate distribution.

Proposition 13 *We suggest a Laplace approximation of this conditional distribution as described in [Rue et al., 2009] which is the multivariate Gaussian distribution with mean* $\hat{\psi}_i$ *and variance-covariance*

$$\Gamma_i = \left(- \mathrm{H}_{\psi}^{\log p}(\hat{\psi}_i) + \Omega^{-1} \right)^{-1} , \tag{6.4.2}$$

where $\mathrm{H}_{\psi}^{\log p}(\hat{\psi}_i) \in \mathbb{R}^{p \times p}$ *is the Hessian of* $\log\left(p_i(y_i|\psi_i)\right)$ *evaluated at* $\hat{\psi}_i$.

Mathematical details for computing this proposal are postponed to Appendix 6.7.2. We use this multivariate Gaussian distribution as a proposal in our IMH algorithm introduced in the next section, for both continuous and noncontinuous data models.

Remark 6.2 *Note that the resulting proposal distribution is based on the assumption that, in model (6.2.2), the random effects* η_i *are normally distributed. When this assumption does not hold, our method exploits the same Gaussian proposal, where the variance* Ω *in (6.4.2) is calculated explicitely. Consider the following example: the random effects* η_i *in (6.2.2) are no longer distributed according to a multivariate Gaussian distribution but a multivariate Student distribution with* d *degrees of freedom, zero mean and a prior shape matrix* ξ *such that* $\eta_i \sim t_d(0, \xi)$. *Then the vector of parameters of the model is* $\boldsymbol{\theta} = (\psi_{\text{pop}}, \Omega, \sigma^2)$ *where* $\Omega = \frac{d}{d-2}\xi$ *is the prior covariance matrix. In that case, our method uses the Independent proposal in Proposition 13 and computes the MH acceptance ratio (6.3.2) with the corresponding multivariate Student density* $p_i(\psi_i)$.

We shall now see another method to derive a Gaussian proposal distribution in the specific case of continuous data models (see (6.2.3)).

6.4.2 Nonlinear continuous data models

When the model is described by (6.2.3), the approximation of the target distribution can be done twofold: either by using the Laplace approximation, as explained above, or by linearizing the structural model f for any individual i of the population. using (6.2.3) and (6.4.1), the MAP estimate can thus be derived as:

$$\hat{\psi}_i = \arg\min_{\psi_i \in \mathbb{R}^p} \left(\frac{1}{\sigma^2} \|y_i - \mathsf{f}(\psi_i)\|^2 + (\psi_i - m_i)' \Omega^{-1} (\psi_i - m_i) \right) . \tag{6.4.3}$$

where $\mathsf{f}(\psi_i)$ is defined by (6.2.4) and A' is the transpose of the matrix A.

We linearize the structural model f around the MAP estimate $\hat{\psi}_i$:

$$\mathsf{f}(\psi_i) \approx \mathsf{f}(\hat{\psi}_i) + \mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)(\psi_i - \hat{\psi}_i) , \tag{6.4.4}$$

where $\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i) \in \mathbb{R}^{n_i \times p}$ is the Jacobian of f evaluated at $\hat{\psi}_i$. Defining $z_i := y_i - \mathsf{f}(\hat{\psi}_i) + \mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)\hat{\psi}_i$, this expansion yields the following linear model:

$$z_i = \mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)\psi_i + \varepsilon_i . \tag{6.4.5}$$

We can directly use the definition of the conditional distribution under a linear model (see (6.7.11) in Appendix 6.7.3) to get an expression of the conditional covariance Γ_i of ψ_i given z_i under (6.4.5):

$$\Gamma_i = \left(\frac{\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)' \, \mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)}{\sigma^2} + \Omega^{-1} \right)^{-1} . \tag{6.4.6}$$

Using (6.4.3) and the definition of the conditional distribution under a linear model we obtain that $\mu_i = \hat{\psi}_i$ (See Appendix 6.7.4 for details). We note that the mode of the conditional distribution of ψ_i in the nonlinear model (6.2.3) is also the mode and the mean of the conditional distribution of ψ_i in the linear model (6.4.5).

Proposition 14 *In the case of continuous data models, we propose to use the multivariate Gaussian distribution, with mean $\hat{\psi}_i$ and variance-covariance matrix Γ_i defined by (6.4.6) as a proposal for an independent MH algorithm avoiding the computation of an Hessian matrix.*

We can note that linearizing the structural model is equivalent to using the Laplace approximation with the expected information matrix. Indeed:

$$\mathbb{E}_{y_i|\hat{\psi}_i} \left(-\mathrm{H}_\psi^{\log p}(\hat{\psi}_i) \right) = \frac{\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)' \, \mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)}{\sigma^2} . \tag{6.4.7}$$

Remark 6.3 *When the model is linear, the probability of accepting a candidate generated*

with this proposal is equal to 1.

Remark 6.4 *If we consider a more general error model, $\varepsilon_i \sim \mathcal{N}(0, \Sigma(t_i, \psi_i))$ that may depend on the individual parameters ψ_i and the observation times t_i, then the conditional variance-covariance matrix reads:*

$$\Gamma_i = \left(\mathrm{J}^{\mathsf{f}}_{\psi}(\hat{\psi}_i)' \Sigma(t_i, \hat{\psi}_i)^{-1} \, \mathrm{J}^{\mathsf{f}}_{\psi}(\hat{\psi}_i) + \Omega^{-1} \right)^{-1} . \tag{6.4.8}$$

Remark 6.5 *In the model (6.7.1), the transformed variable $\phi_i = u(\psi_i)$ follows a normal distribution. Then a candidate ϕ_i^c is drawn from the multivariate Gaussian proposal with parameters:*

$$\mu_i = \hat{\phi}_i \, , \tag{6.4.9}$$

$$\Gamma_i = \left(\frac{\mathrm{J}^{\mathsf{f}}_{\psi}(u^{-1}(\hat{\phi}_i))' \, \mathrm{J}^{\mathsf{f}}_{\psi}(u^{-1}(\hat{\phi}_i))}{\sigma^2} + \Omega^{-1} \right)^{-1} , \tag{6.4.10}$$

where $\hat{\phi}_i = \arg\max_{\phi_i \in \mathbb{R}^p} p_i(\phi_i | y_i)$ and finally the candidate vector of individual parameters is set to $\psi_i^c = u^{-1}(\phi_i^c)$

These approximations of the conditional distribution $p_i(\psi_i|y_i)$ lead to our nlme-IMH algorithm, an Independent Metropolis-Hastings (IMH) algorithm for NLME models. For all individuals $i \in [\![1, n]\!]$, the algorithm is defined as:

Algorithm 6.2 The nlme-IMH algorithm

Initialization: Initialize the chain sampling $\psi_i^{(0)}$ from some initial distribution ξ_i .
Iteration t: Given the current state of the chain $\psi_i^{(t-1)}$:

1. Compute the MAP estimate:

$$\hat{\psi}_i^{(t)} = \arg\max_{\psi_i \in \mathbb{R}^p} p_i(\psi_i | y_i) \, . \tag{6.4.11}$$

2. Compute the covariance matrix $\Gamma_i^{(t)}$ using either (6.4.2) or (6.4.6).
3. Sample a candidate ψ_i^c from a the independent proposal $\mathcal{N}(\hat{\psi}_i^{(t)}, \Gamma_i^{(t)})$ denoted $q_i(\cdot|\hat{\psi}_i^{(t)})$.
4. Compute the MH ratio:

$$\alpha(\psi_i^{(t)}, \psi_i^c) = \frac{p_i(\psi_i^c|y_i)}{p_i(\psi_i^{(t)}|y_i)} \frac{q_i(\hat{\psi}_i^{(t)}|\psi_i^c)}{q_i(\psi_i^c|\hat{\psi}_i^{(t)})} \, . \tag{6.4.12}$$

5. Set $\psi_i^{(t)} = \psi_i^c$ with probability $\min(1, \alpha(\psi_i^c, \psi_i^{(t-1)}))$ (otherwise, keep $\psi_i^{(t)} = \psi_i^{(t-1)}$).

This method shares some similiarities with [Titsias and Papaspiliopoulos, 2018] that suggests to perform a Taylor expansion of $p_i(y_i|\psi_i)$ around the current state of the chain, leaving $p_i(\psi_i)$ unchanged.

Remark 6.6 *Although a multivariate Gaussian proposal is used in our presentation of the nlme-IMH, other type of distributions could be adopted. For instance, when the target distribution presents heavy tails, a Student distribution with a well-chosen degree of freedom could improve the performance of the independent sampler. In such case, the parameters of the Gaussian proposal are used to shift and scale the Student proposal distribution and the acceptance rate (6.4.12) needs to be modified accordingly. The numerical applications in Section 5 are performed using a Gaussian proposal but comparisons with a Student proposal distribution are given in Appendix 6.8.1.*

6.4.3 Maximum Likelihood Estimation

The ML estimator defined by (6.3.1) is computed using the Stochastic Approximation of the EM algorithm (SAEM) [Delyon et al., 1999b]. The SAEM algorithm is described as follows:

Algorithm 6.3 The SAEM algorithm

Initialization: $\boldsymbol{\theta}_0$, an initial parameter estimate and M, the number of MCMC iterations.
Iteration k: given the current model parameter estimate $\boldsymbol{\theta}^{(k-1)}$:

1. **Simulation step:** For $i \in [\![1, n]\!]$, draw a vector of individual parameters $\psi_i^{(k)}$ resulting from M iterations of the transition kernel $\Pi_i^{(k)}$, starting from $\psi_i^{(k-1)}$, which admits as unique limiting distribution the conditional distribution $\mathtt{p}_i(\psi_i|y_i;\theta_{k-1})$.
2. **Stochastic approximation step:** update the approximation of the conditional expectation $\mathbb{E}\left[\log \mathtt{p}(y,\psi;\boldsymbol{\theta})|y,\boldsymbol{\theta}^{(k-1)}\right]$:

$$Q_k(\boldsymbol{\theta}) = Q_{k-1}(\boldsymbol{\theta}) + \gamma_k \left(\sum_{i=1}^{n} \log f_i(y_i, \psi_i^{(k)}; \boldsymbol{\theta}) - Q_{k-1}(\boldsymbol{\theta}) \right) , \tag{6.4.13}$$

   where $\{\gamma_k\}_{k>0}$ is a sequence of decreasing stepsizes with $\gamma_1 = 1$.
3. **Maximisation step:** Update the model parameter estimate:

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta}\in\mathbb{R}^d} Q_k(\boldsymbol{\theta}) . \tag{6.4.14}$$

The SAEM algorithm is implemented in most sofware tools for NLME models and its convergence is studied in [Allassonniere and Kuhn, 2013, Delyon et al., 1999b, Kuhn and Lavielle, 2004]. The practical performances of SAEM are closely linked to the settings of SAEM. In particular, the choice of the transition kernel Π plays a key role. The transition kernel Π is directly defined by the proposal(s) used for the MH algorithm.

We propose a fast version of the SAEM algorithm using our resulting independent proposal distribution called the f-SAEM. The simulation step of the f-SAEM is achieved using the nlme-IMH algorithm (see algorithm 6.2) for all individuals $i \in [\![1, n]\!]$ and the next steps remain unchanged. In practice, the number of transitions M is small since the convergence

of the SAEM does not require the convergence of the MCMC at each iteration [Kuhn and Lavielle, 2004]. In the sequel, we carry out numerous numerical experiments to compare our nlme-IMH algorithm to state-of-the-art samplers and assess its relevance in a MLE algorithm such as the SAEM.

6.5 Application to Pharmacology

6.5.1 A pharmacokinetic example

6.5.1.1 Data and model

32 healthy volunteers received a 1.5 mg/kg single oral dose of warfarin, an anticoagulant normally used in the prevention of thrombosis [O'Reilly and Aggeler, 1968]. Figure 6.1 shows the warfarin plasmatic concentration measured at different times for these patients (the single dose was given at time 0 for all the patients).

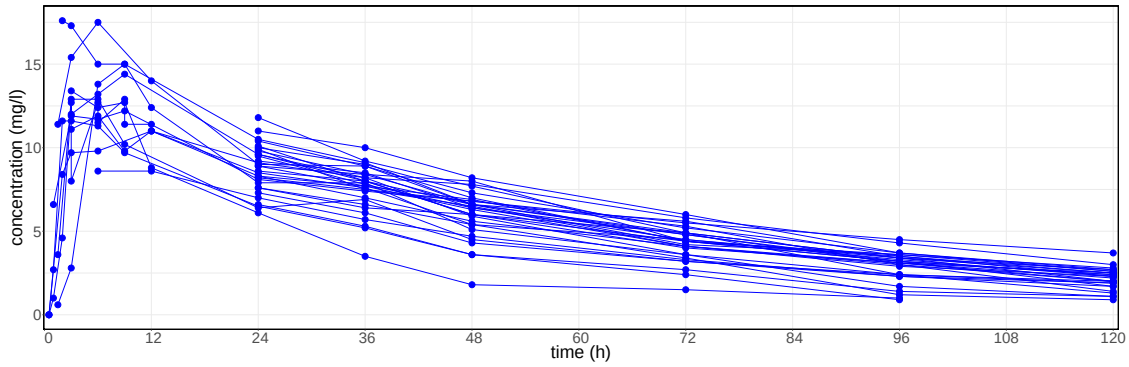

Figure 6.1 – Warfarin concentration (mg/l) over time (h) for 32 subjects

We consider a one-compartment pharmacokinetics (PK) model for oral administration, assuming first-order absorption and linear elimination processes:

$$\mathsf{f}(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t}) \;, \tag{6.5.1}$$

where ka is the absorption rate constant, V the volume of distribution , k the elimination rate constant, and D the dose of drug administered. Here, ka, V and k are PK parameters that can change from one individual to another. Let $\psi_i = (ka_i, V_i, k_i)$ be the vector of individual PK parameters for individual i. The model for the j-th measured concentration, noted y_{ij}, for individual i writes:

$$y_{ij} = \mathsf{f}(t_{ij}, \psi_i) + \varepsilon_{ij} \;. \tag{6.5.2}$$

We assume in this example that the residual errors are independent and normally distributed with mean 0 and variance σ^2. Lognormal distributions are used for the three PK parameters:

$$\log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2) \ , \log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2) \ , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2) \ . \tag{6.5.3}$$

This is a specific instance of the nonlinear mixed effects model for continuous data described in Section 6.2.2. We thus use the multivariate Gaussian proposal whose mean and covariance are defined by (6.7.13) and (6.4.6). In such case the gradient can be explicitly computed. Nevertheless, for the method to be easily extended to any structural model, the gradient is calculated using Automatic Differentiation [Griewank and Walther, 2008] implemented in the R package "Madness" [Pav, 2016].

6.5.1.2 MCMC Convergence Diagnostic

We study in this section the behaviour of the MH algorithm used to sample individual parameters from the conditional distribution $p_i(\psi_i|y_i;\boldsymbol{\theta})$. We consider only one of the 32 individuals for this study and fix $\boldsymbol{\theta}$ close to the ML estimate obtained with the SAEM algorithm, implemented in the SAEMIX R package [Comets et al., 2017]: $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.01$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$.

We run the classical version of MH implemented in the SAEMIX package and for which different transition kernels are used successively at each iteration: independent proposals from the marginal distribution $p_i(\psi_i)$, component-wise random walk and block-wise random walk. We compare it to our proposed algorithm 6.2.

We run 20 000 iterations of these two algorithms and evaluate their convergence by looking at the convergence of the median for the three components of ψ_i. We see Figure 6.2 that, for parameter k_i, the sequences of empirical median obtained with the two algorithms converge to the same value, which is supposed to be the theoretical median of the conditional distribution. It is interesting to note that the empirical median with the nlme-IMH converge very rapidly. This is interesting in the population approach framework because it is mainly the median values of each conditional distribution that are used to infer the population distribution. Autocorrelation plots, Figure 6.2, highlight slower mixing of the RWM whereas samples from the nlme-IMH can be considered independent few iterations after the chain has been initialized. Comparison for all three dimensions of the individual parameter ψ_i using a Student proposal distribution can be found in Appendix 6.8.1.

The Mean Square Jump Distance (MSJD) as well as the Effective Sample Size (ESS) of the two methods are reported in Table 6.5. MSJD is a measure used to diagnose the mixing

of the chain. It is calculated as the mean of the squared euclidean distance between every point and its previous point. Usually, this quantity indicates if the chain is moving enough or getting stuck at some region and the ESS is a quantity that estimates the number of independent samples obtained from a chain. Larger values of those two quantities for our method show greater performance of the sampler in comparison with the RWM.

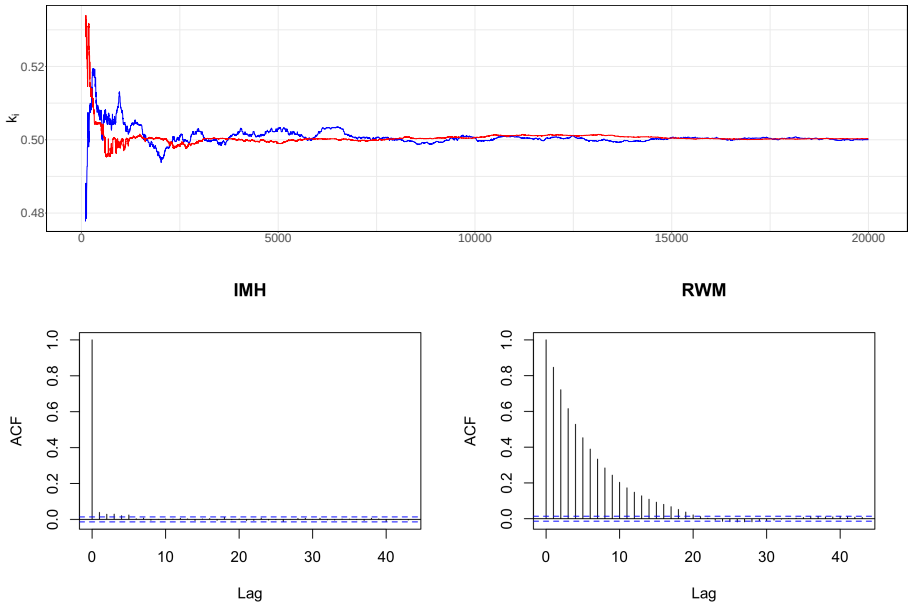

Figure 6.2 – Modelling of the warfarin PK data. Top plot: convergence of the empirical medians of $p_i(k_i|y_i;\boldsymbol{\theta})$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter k_i.

Table 6.1 – MSJD and ESS per dimension.

	MSJD			**ESS**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH	0.061	0.004	0.018	13694	14907	19976

Comparison with state-of-the-art methods: We then compare our new approach to the three following samplers: an independent sampler that uses variational approximation as proposal distribution [de Freitas et al., 2001], the MALA [Roberts and Tweedie, 1996] and the No-U-Turn Sampler [Hoffman and Gelman, 2014].

The same design and settings (dataset, model parameter estimate, individual) as in section 6.5.1.2 are used throughout the following experiments.

Variational MCMC algorithm

The Variational MCMC algorithm [de Freitas et al., 2001] is a MCMC algorithm with independent proposal. The proposal distribution is a multivariate Gaussian distribution whose parameters are obtained by a variational approach that consists in minimising the Kullback Leibler divergence between a multivariate Gaussian distribution $q_i(\psi_i, \delta)$, and the target distribution for a given model parameter estimate $\boldsymbol{\theta}$ noted $p_i(\psi_i|y_i, \boldsymbol{\theta})$. This problem boils down to maximizing the so-called Evidence Lower Bound ELBO($\boldsymbol{\theta}$) defined as:

$$\text{ELBO}(\delta) \triangleq \int q_i(\psi_i, \delta) \left(\log f_i(y_i, \psi_i, \boldsymbol{\theta}) - \log q_i(\psi_i, \delta)\right) \mathrm{d}\psi_i \ . \tag{6.5.4}$$

We use the Automatic Differentiation Variational Inference (ADVI) [Kucukelbir et al., 2015] implemented in RStan (R Package [Stan Development Team, 2018]) to obtain the vector of parameters noted δ_{VI} defined as:

$$\delta_{VI} \triangleq \underset{\delta \in \mathbb{R}^p \times \mathbb{R}^{p \times p}}{\arg\max} \ \text{ELBO}(\delta) \ .$$

The algorithm stops when the variation of the median of the objective function falls below the 1% threshold. The means and standard deviations of our nlme-IMH and the Variational MCMC proposals compare with the posterior mean (calculated using the NUTS [Hoffman and Gelman, 2014]) as follows:

Table 6.2 – Means and standard deviations.

	Means			**Stds**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
Variational proposal	0.90	7.93	0.48	0.14	0.03	0.07
Laplace proposal	0.88	7.93	0.52	0.18	0.04	0.09
NUTS (ground truth)	0.91	7.93	0.51	0.18	0.05	0.09

We observe that the mean of the variational approximation is slightly shifted from the estimated posterior mode (see table 6.2 for comparison) whereas a considerable difference lies in the numerical value of the covariance matrix obtained with ADVI. The empirical standard deviation of the Variational MCMC proposal is much smaller than our new proposal defined by (6.4.6) (see table 6.2), which slows down the MCMC convergence.

Figure 6.3 shows the proposals marginals and the marginal posterior distribution for the individual parameters k_i and V_i. Biplot of the samples drawn from the two multivariate Gaussian proposals (our independent proposal and the variational MCMC proposal) as well as samples drawn from the posterior distribution (using the NUTS) are also presented in this figure. We conclude that both marginal and bivariate posterior distributions are better approximated by our independent proposal than the one resulting from a KL divergence optimization.

Besides similar marginal variances, both our independent proposal and the true posterior share a strong anisotropic nature, confirmed by the similar correlation values of table 6.6 (see Appendix 6.8.1). Same characteristics are observed for the other parameters. Those highlighted properties leads to a better performance of the nlme-IMH versus the ADVI sampler as reflected in Figure 6.4. Larger jumps of the chain and bigger ESS show how effective the nlme-IMH is compared to the ADVI (see Table 6.3).

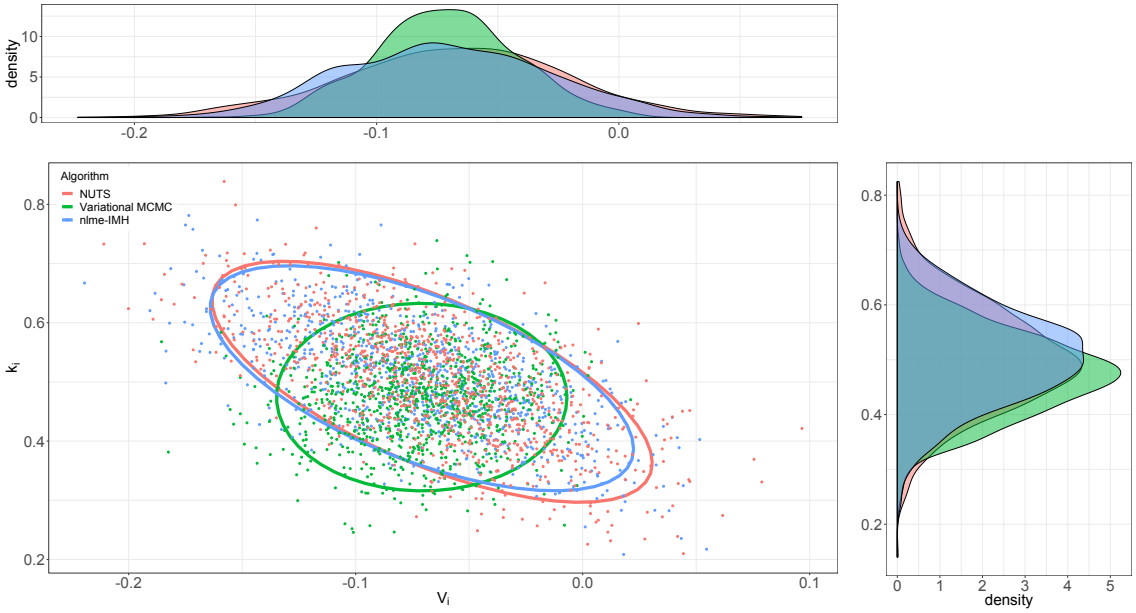

Figure 6.3 – Modelling of the warfarin PK data: Comparison between the proposals of the nlme-IMH (blue), the Variational MCMC (green) and the empirical target distribution sampled using the NUTS (red). Marginals and biplots of the conditional distributions $k_i|y_i$ and $V_i|y_i$ for a single individual. Ellipses containing 90% of the data points are represented on the main plot.

Metropolis Adjusted Langevin Algorithm (MALA) and No-U-Turn Sampler (NUTS)

We now compare our method to the MALA, which proposal is defined by (6.3.3). The gradient of the log posterior distribution $\nabla_{\psi_i} \log p_i(\psi_i^{(k)}|y_i)$ is also calculated by Automatic Differentiation. In this numerical example, the MALA has been initialized at the MAP and the stepsize ($\gamma = 10^{-2}$) is tuned such that the acceptance rate of 0.57 is reached [Roberts and Rosenthal, 1997].

We also compare the implementation of NUTS [Carpenter et al., 2017, Hoffman and Gelman, 2014] in the RStan package to our method in Figure 6.4. Figure 6.4 highlights good convergence of a well-tuned MALA and the NUTS. nlme-IMH and NUTS mixing properties, from autocorrelation plots in Figure 6.4 seem to be similar and much better than all of the other methods. Table 6.3 presents a benchmark of those methods regarding MSJD

and ESS. Both nlme-IMH and NUTS have better performances here. For parameters ka and V, the ESS of the NUTS, presented as a gold standard sampler for this king of problem, are slightly higher than our proposed method.

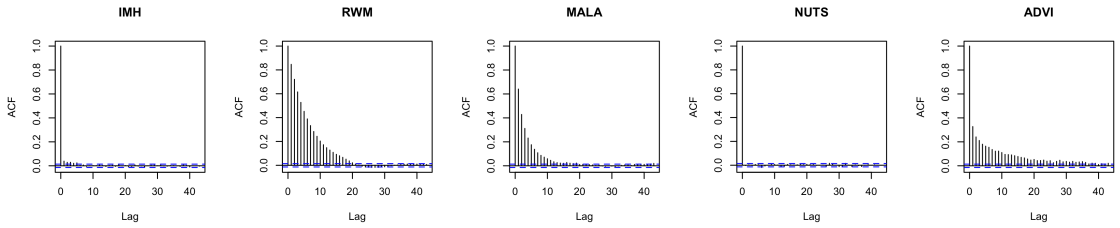

Figure 6.4 – Modelling of the warfarin PK data: Autocorrelation plots of the MCMC samplers for parameter k_i.

Table 6.3 – MSJD and ESS per dimension.

	MSJD			**ESS**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH	0.061	0.004	0.018	13694	14907	19976
MALA	0.024	0.002	0.006	3458	3786	3688
NUTS	0.063	0.004	0.018	18684	19327	19083
ADVI	0.037	0.002	0.010	2499	1944	2649

In practice, those three methods imply tuning phases that are computationally involved , warming up the chain and a careful initialisation whereas our independent sampler is automatic and fast to implement. Investigating the asymptotic convergence behavior of those methods highlights the competitive properties of our IMH algorithm to sample from the target distribution.

Since our goal is to embed those samplers into a MLE algorithm such as the SAEM, we shall now study how they behave in the very first iterations of the MCMC procedure. Recall that the SAEM requires only few iterations of MCMC sampling under the current model parameter estimate. We present this non asymptotic study in the following section.

6.5.1.3 Comparison of the chains for the first 500 iterations

We produce 100 independent runs of the RWM, the nlme-IMH, the MALA and the NUTS for 500 iterations. The boxplots of the samples drawn at a given iteration threshold (three different thresholds are used) are presented Figure 6.5 against the ground truth for the parameter **ka**. The ground truth has been calculated by running the NUTS for 100 000 iterations.

For the three numbers of iteration (5,20,500) considered in Figure 6.5, the median of the nlme-IMH and NUTS samples are closer to the ground truth. Figure 6.5 also highlights

that all those methods succeed in sampling from the whole distribution after 500 iterations.

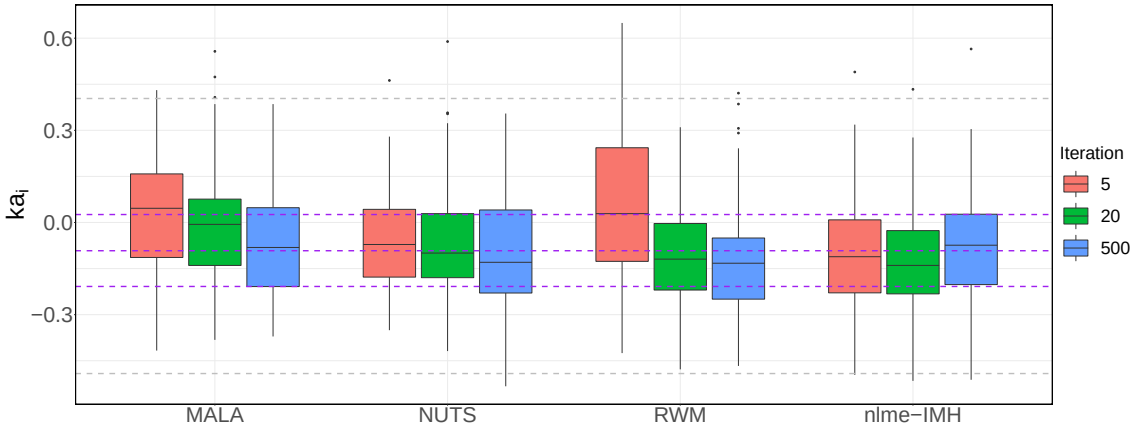

Figure 6.5 – Modelling of the warfarin PK data: Boxplots for the RWM, the nlme-IMH, the MALA and the NUTS algorithm, averaged over 100 independent runs. The groundtruth median, 0.25 and 0.75 percentiles are plotted as a dashed purple line and its maximum and minimum as a dashed grey line.

We now use the RWM, the nlme-IMH and the MALA in the SAEM algorithm and observe the convergence of the resulting sequences of parameters.

6.5.1.4 Maximum likelihood estimation

We use the SAEM algorithm to estimate the population PK parameters ka_{pop}, V_{pop} and k_{pop}, the standard deviations of the random effects ω_{k_a}, ω_V and ω_k and the residual variance σ^2.

The stepsize γ_k is set to 1 during the first 100 iterations and then decreases as $1/k^a$ where $a = 0.7$ during the next 100 iterations.

Here we compare the standard SAEM algorithm, as implemented in the SAEMIX R package, with the f-SAEM algorithm and the SAEM using the MALA sampler. In this example, the nlme-IMH and the MALA are only used during the first 20 iterations of the SAEM. The standard MH algorithm is then used.

Figure 6.6 shows the estimates of V_{pop} and ω_V computed at each iteration of these three variants of SAEM and starting from three different initial values. First of all, we notice that, whatever the initialisation and the sampling algorithm used, all the runs converge towards the maximum likelihood estimate. It is then very clear that the f-SAEM converges faster than the standard algorithm. The SAEM using the MALA algorithm for sampling from the individual conditional distribution presents a similar convergence behavior as the reference.

We can conclude, for this example, that sampling around the MAP of each individual

conditional distribution is the key to a fast convergence of the SAEM during the first iterations.

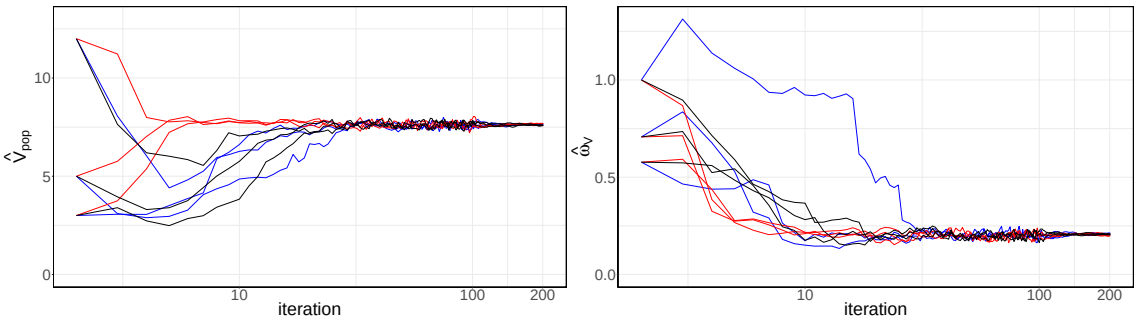

Figure 6.6 – Estimation of the population PK parameters for the warfarin data: convergence of the sequences of estimates $\{\hat{V}_{\text{pop}}^{(k)}\}_{1\leq k\leq 200}$ and $\{\hat{\omega}_V^{(k)}\}_{1\leq k\leq 200}$ obtained with SAEM and three different initial values using the reference MH algorithm (blue), the f-SAEM (red) and the SAEM using the MALA sampler (black).

6.5.1.5 Monte Carlo study

We conduct a Monte Carlo study to confirm the properties of the f-SAEM algorithm for computing the ML estimates.

$M = 50$ datasets have been simulated using the PK model previously used for fitting the warfarin PK data with the following parameter values: $ka_{\text{pop}} = 1$, $V_{\text{pop}} = 8$, $k_{\text{pop}} = 0.1$, $\omega_{ka} = 0.5$, $\omega_V = 0.2$, $\omega_k = 0.3$ and $\sigma^2 = 0.5$. The same original design with $N = 32$ patients and a total number of 251 PK measurements were used for all the simulated datasets. Since all the simulated data are different, the value of the ML estimator varies from one simulation to another. If we run K iterations of SAEM, the last element of the sequence $\{(\boldsymbol{\theta}^{(k)})^{(m)}\}_{1\leq k\leq K}$ is the estimate obtained from the m-th simulated dataset. To investigate how fast $((\boldsymbol{\theta}^{(k)})^{(m)}, 1 \leq k \leq K)$ converges to $(\boldsymbol{\theta}^{(k)})^{(m)}$ we study how fast $\{(\boldsymbol{\theta}^{(k)})^{(m)} - (\boldsymbol{\theta}^{(K)})^{(m)}\}_{1\leq k\leq K}$ goes to 0. For a given sequence of estimates, we can then define, at each iteration k and for each component ℓ of the parameter, the mean square distance over the replicates

$$E^{(k)}(\ell) = \frac{1}{M}\sum_{m=1}^{M}\left((\boldsymbol{\theta}^{(k)})^{(m)}(\ell) - (\boldsymbol{\theta}^{(k)})^{(m)}(\ell)\right)^2 . \tag{6.5.5}$$

Figure 6.7 shows using the new proposal leads to a much faster convergence towards the maximum likelihood estimate. Less than 10 iterations are required to converge with the f-SAEM on this example, instead of 50 with the original version. It should also be noted that the distance decreases monotonically. The sequence of estimates approaches the target at each iteration, compared to the standard algorithm which makes twists and turns before converging.

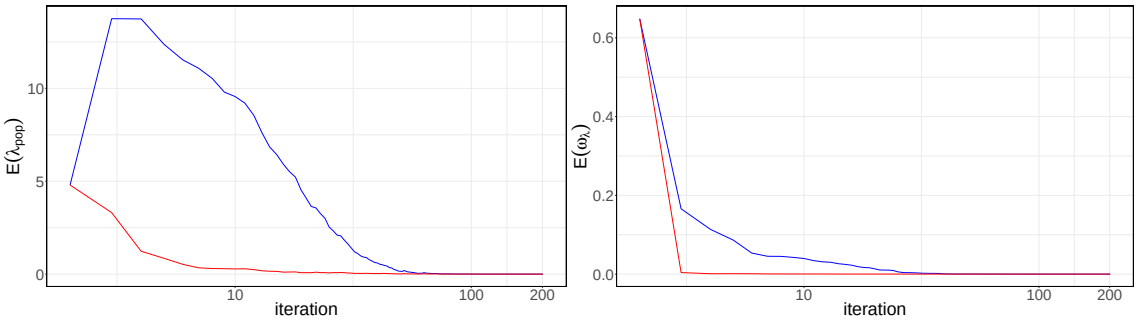

Figure 6.7 – Convergence of the sequences of mean square distances $(E^{(k)}(V_{\text{pop}}), 1 \leq k \leq 200)$ and $(E^{(k)}(\omega_V), 1 \leq k \leq 200)$ for V_{pop} and ω_V obtained with SAEM on $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the f-SAEM (red).

6.5.2 Time-to-event Data Model

6.5.2.1 The model

In this section, we consider a Weibull model for time-to-event data [Lavielle, 2014, Zhang, 2016]. For individual i, the hazard function of this model is:

$$h(t, \psi_i) = \frac{\beta_i}{\lambda_i} \left(\frac{t}{\lambda_i} \right)^{\beta_i - 1} . \tag{6.5.6}$$

Here, the vector of individual parameters is $\psi_i = (\lambda_i, \beta_i)$. These two parameters are assumed to be independent and lognormally distributed:

$$\log(\lambda_i) \sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2), \log(\beta_i) \sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2) . \tag{6.5.7}$$

Then, the vector of population parameters is $\boldsymbol{\theta} = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda, \omega_\beta)$.

Repeated events were generated, for $N = 100$ individuals, using the Weibull model (8.3.3) with $\lambda_{\text{pop}} = 10$, $\omega_\lambda = 0.3$, $\beta_{\text{pop}} = 3$ and $\omega_\beta = 0.3$ and assuming a right censoring time $\tau_c = 20$.

6.5.2.2 MCMC Convergence Diagnostic

Similarly to the previous section, we start by looking at the behaviour of the MCMC procedure used for sampling from the conditional distribution $p_i(\psi_i | y_i; \boldsymbol{\theta})$ for a given individual i and assuming that $\boldsymbol{\theta}$ is known. We use the generating model parameter in these experiments ($\boldsymbol{\theta} = (\lambda_{\text{pop}} = 10, \beta_{\text{pop}} = 3, \omega_\lambda = 0.3, \omega_\beta = 0.3)$).

We run 12 000 iterations of the reference MH algorithm the nlme-IMH to estimate the median of the posterior distribution of λ_i. We see Figure 6.8 that the sequences of empirical medians obtained with the two procedures converge to the same value but the

new algorithm converges faster than the standard MH algorithm. Autocorrelation plots, Figure 6.8, are also significantly showing the advantage of the new sampler as the chain obtained with the nlme-IMH is mixing almost ten times faster than the reference sampler.

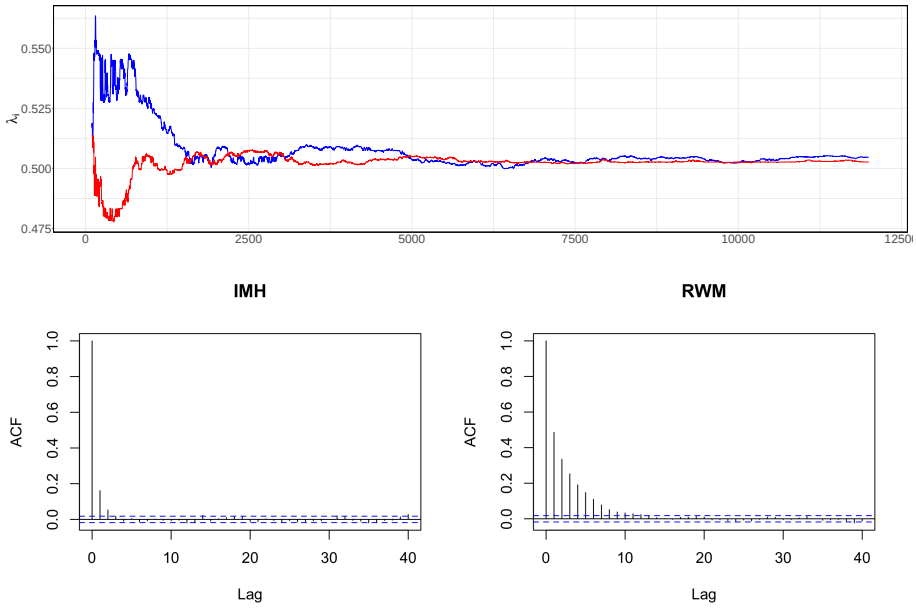

Figure 6.8 – Time-to-event data modelling. Top plot: convergence of the empirical medians of $p_i(\lambda_i|y_i;\boldsymbol{\theta})$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter λ_i.

Table 6.4 – MSJD and ESS per dimension.

	MSJD		**ESS**	
	λ_i	β_i	λ_i	β_i
RWM	0.055	0.093	3061	1115
nlme-IMH	0.095	0.467	8759	8417

Plots for the other parameter can be found in Appendix 6.8.2. Comparisons with state-of-the-art methods were conducted as in the previous section. These comparisons led us to the same remarks as those made for the previous continuous data model both on the asymptotic and non asymptotic regimes.

6.5.2.3 Maximum likelihood estimation of the population parameters

We run the standard SAEM algorithm implemented in the SAEMIX package (extension of this package for noncontinuous data models is available on GitHub: `https://github.com/belhal/saemix`) and the f-SAEM on the generated dataset.

Figure 6.9 shows the estimates of λ_{pop} and ω_λ computed at each iteration of the two versions of the SAEM and starting from three different initial values. The same behaviour

is observed as in the continuous case: regardless the initial values and the algorithm, all the runs converge to the same solution but convergence is much faster with the proposed method. The same comment applies for the two other parameters β_{pop} and ω_β.

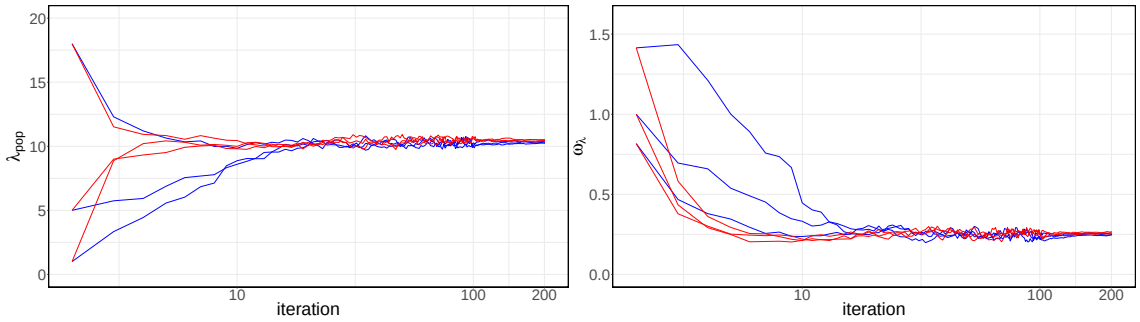

Figure 6.9 – Population parameter estimation in time-to-event-data models: convergence of the sequences of estimates $\{\hat{\lambda}_{\text{pop}}^{(k)}\}_{1\leq k\leq 200}$ and $\{\hat{\omega}_\lambda^{(k)}\}_{1\leq k\leq 200}$ obtained with SAEM and three different initial values using the reference MH algorithm (blue) and the f-SAEM (red).

6.5.2.4 Monte Carlo study

We now conduct a Monte Carlo study in order to confirm the good properties of the new version of the SAEM algorithm for estimating the population parameters of a time-to-event data model. $M = 50$ synthetic datasets are generated using the same design as above. Figure 6.10 shows the convergence of the mean square distances defined in (6.5.5) for λ_{pop} and ω_λ. All these distances converge monotonically to 0 which means that both algorithms properly converge to the maximum likelihood estimate, but very few iterations are required with the new version to converge while about thirty iterations are needed with the SAEM.

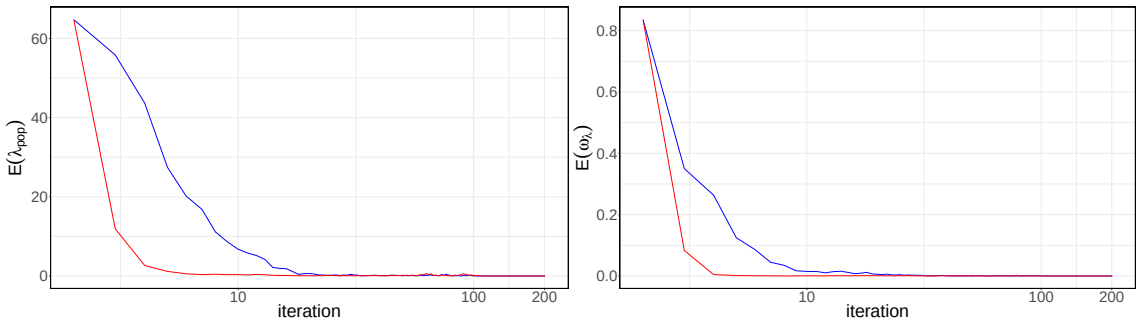

Figure 6.10 – Convergence of the sequences of mean square distances $(E^{(k)}(\lambda_{\text{pop}}), 1 \leq k \leq 200)$ and $(E^{(k)}(\omega_\lambda), 1 \leq k \leq 200)$ for λ_{pop} and ω_λ obtained with SAEM from $M = 50$ synthetic datasets using the reference MH algorithm (blue) and the f-SAEM (red).

6.6 Conclusion

We present in this article an independent Metropolis-Hastings procedure for sampling random effects from their conditional distributions and a fast MLE algorithm, called the f-SAEM, in nonlinear mixed effects models.

The idea of the method is to approximate each individual conditional distribution by a multivariate normal distribution. A Laplace approximation makes it possible to consider any type of data, but we have shown that, in the case of continuous data, this approximation is equivalent to linearizing the structural model around the conditional mode of the random effects.

The numerical experiments demonstrate that the proposed nlme-IMH sampler converges faster to the target distribution than a standard random walk Metropolis. This practical behaviour is partly explained by the fact that the conditional mode of the random effects in the linearized model coincides with the conditional mode of the random effects in the original model. The proposal distribution is therefore a normal distribution centered around this MAP. On the other hand, the dependency structure in the conditional distribution of the random effects is well approximated by the covariance structure of the Gaussian proposal. So far, we have mainly applied our method to standard problems encountered in pharmacometrics and for which the number of random effects remains small. It can nevertheless be interesting to see how this method behaves in higher dimension and compare it with methods adapted to such situations such as MALA or HMC. Lastly, we have shown that this new IMH algorithm can easily be embedded in the SAEM algorithm for maximum likelihood estimation of the population parameters. Our numerical studies have shown empirically that the new transition kernel is effective in the very first iterations. It is of great interest to determine automatically and in an adaptive way an optimal scheme of kernel transitions combining this new proposal with the block-wise random walk Metropolis.

Appendices to Fast Stochastic Approximation of the EM

6.7 Mathematical Details

6.7.1 Extensions of model (6.2.2)

Several extensions of model (6.2.2) are also possible. We can assume for instance that the transformed individual parameters are normally distributed:

$$u(\psi_i) = u(\psi_{\text{pop}}) + \eta_i \ , \tag{6.7.1}$$

where u is a strictly monotonic transformation applied on the individual parameters ψ_i. Examples of such transformation are the logarithmic function (in which case the components of ψ_i are log-normally distributed), the logit and the probit transformations [Lavielle, 2014]. In the following, we either use the original parameter ψ_i or the Gaussian transformed parameter $u(\psi_i)$.

Another extension of model (6.2.2) consists in introducing individual covariates in order to explain part of the inter-individual variability:

$$u(\psi_i) = u(\psi_{\text{pop}}) + C_i\beta + \eta_i \ , \tag{6.7.2}$$

where C_i is a matrix of individual covariates. Here, the fixed effects are the vector of coefficients β and the vector of typical parameters ψ_{pop}.

6.7.2 Calculus of the proposal in the noncontinuous case

Laplace approximation (see [Migon et al., 2014]) consists in approximating an integral of the form

$$I := \int \mathrm{e}^{v(x)} \, \mathrm{d}x \ , \tag{6.7.3}$$

where v is at least twice differentiable.

The following second order Taylor expansion of the function v around a point x_0

$$v(x) \approx v(x_0) + \nabla v(x_0)(x - x_0) + \frac{1}{2}(x - x0)\nabla^2 v(x_0)(x - x0) \ , \tag{6.7.4}$$

provides an approximation of the integral I (consider a multivariate Gaussian probability distribution function which integral sums to 1):

$$I \approx \mathrm{e}^{v(x_0)} \sqrt{\frac{(2\pi)^p}{|-\nabla^2 v(x_0)|}} \exp\left\{-\frac{1}{2}\nabla v(x_0)'\nabla^2 v(x_0)^{-1}\nabla v(x_0)\right\} \ . \tag{6.7.5}$$

In our context, we can write the marginal pdf $p_i(y_i)$ that we aim to approximate as

$$p_i(y_i) = \int f_i(y_i, \psi_i)\mathrm{d}\psi_i = \int \mathrm{e}^{\log f_i(y_i,\psi_i)} \, \mathrm{d}\psi_i \ . \tag{6.7.6}$$

Then, let

$$v(\psi_i) := \log f_i(y_i, \psi_i) = \log p_i(y_i|\psi_i) + \log p_i(\psi_i) \ , \tag{6.7.7}$$

and compute its Taylor expansion around the MAP $\hat{\psi}_i$. We have by definition that

$$\nabla \log p_i(y_i, \hat{\psi}_i) = 0 \ ,$$

which leads to the following Laplace approximation of $\log p_i(y_i)$:

$$-2\log p_i(y_i) \approx -p\log 2\pi - 2\log p_i(y_i, \hat{\psi}_i) + \log\left(\left|-\nabla^2 \log p_i(y_i, \hat{\psi}_i)\right|\right) \ .$$

We thus obtain the following approximation of the logarithm of the conditional pdf of ψ_i given y_i evaluated at $\hat{\psi}_i$:

$$\log p_i(\hat{\psi}_i|y_i) \approx -\frac{p}{2}\log 2\pi - \frac{1}{2}\log\left(\left|-\nabla^2 \log p_i(y_i, \hat{\psi}_i)\right|\right) \ ,$$

which is precisely the log-pdf of a multivariate Gaussian distribution with mean $\hat{\psi}_i$ and variance-covariance $-\nabla^2 \log p_i(y_i, \hat{\psi}_i)^{-1}$ with:

$$\nabla^2 \log p_i(y_i, \hat{\psi}_i) = \nabla^2 \log p_i(y_i|\hat{\psi}_i) + \nabla^2 \log p_i(\hat{\psi}_i) \tag{6.7.8}$$

$$= \mathrm{H}_{\psi}^{\log p}(\hat{\psi}_i) + \Omega^{-1} \ . \tag{6.7.9}$$

6.7.3 Linear continuous data models

Let $y_i = (y_{i,1}, \ldots, y_{i,n_i})'$ and $\varepsilon_i = (\varepsilon_{i,1}, \ldots, \varepsilon_{i,n_i})'$. Assume a linear relationship between the observations y_i and the vector of individual parameters ψ_i:

$$y_i = A_i\psi_i + \varepsilon_i \ , \tag{6.7.10}$$

where $A_i \in \mathbb{R}^{n_i \times p}$ is the design matrix for individual i, ψ_i is normally distributed with mean $m_i \in \mathbb{R}^p$ and covariance $\Omega \in \mathbb{R}^{p\times p}$. Then, the conditional distribution of ψ_i given y_i is a normal distribution with mean μ_i and variance-covariance matrix Γ_i defined as:

$$\mu_i = \Gamma_i \left(\frac{A_i' y_i}{\sigma^2} + \Omega^{-1} m_i \right) \quad \text{where} \quad \Gamma_i = \left(\frac{A_i' A_i}{\sigma^2} + \Omega^{-1} \right)^{-1} \tag{6.7.11}$$

Here, μ_i is the mode of the conditional distribution of ψ_i, known as the Maximum A Posteriori (MAP) estimate, or the Empirical Bayes Estimate (EBE) of ψ_i.

6.7.4 Conditional mode under the linearised model

Using (6.4.3), $\hat{\psi}_i$ satisfies:

$$-\frac{\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)'}{\sigma^2}\left(y_i - \mathsf{f}(\hat{\psi}_i)\right) + \Omega^{-1}(\hat{\psi}_i - m_i) = 0 \ , \tag{6.7.12}$$

which leads to the definition of the conditional mean μ_i of ψ_i given z_i, under the linearized model, by:

$$\begin{aligned}
\mu_i &= \Gamma_i \frac{\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)'}{\sigma^2}\left(y_i - \mathsf{f}(\hat{\psi}_i) + \mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)\hat{\psi}_i + \Omega^{-1} m_i\right) && (6.7.13)\\
&= \Gamma_i \left(\Omega^{-1}(\hat{\psi}_i - m_i) + \frac{\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)'\,\mathrm{J}_\psi^{\mathsf{f}}(\hat{\psi}_i)}{\sigma^2}\hat{\psi}_i + \Omega^{-1} m_i \right) && (6.7.14)\\
&= \Gamma_i \Gamma_i^{-1} \hat{\psi}_i = \hat{\psi}_i \ . && (6.7.15)
\end{aligned}$$

6.8 Supplementary Experiments

6.8.1 A pharmacokinetic example

Figures 6.11 and 6.12 highlight the performances of the RWM, the nlme-IMH using a Gaussian proposal distribution and a Student proposal. At iteration (t) of the MH algorithm, samples from the Student proposal distribution are obtained using the same parameters obtained in Proposal 14 as follows:

Table 6.5 – MSJD and ESS per dimension.

	MSJD			**ESS**		
	ka_i	V_i	k_i	ka_i	V_i	k_i
RWM	0.009	0.002	0.006	1728	3414	3784
nlme-IMH (Gaussian)	0.061	0.004	0.018	13694	14907	19976
nlme-IMH (Student)	0.063	0.004	0.018	14907	19946	19856

- Student samples $S_i^{(t)}$ are drawn from a student distribution with degree of freedom $k = 3$: $S_i^{(t)} \sim t(k)$
- Individual parameters $\psi_i^{(t)}$ are obtained using the mean and the covariance defined in Proposal 14 to shift and scale the obtained samples: $\psi_i^{(t)} = \hat{\psi}_i^{(t)} + S_i^{(t)}.\Gamma_i^{(t)}$

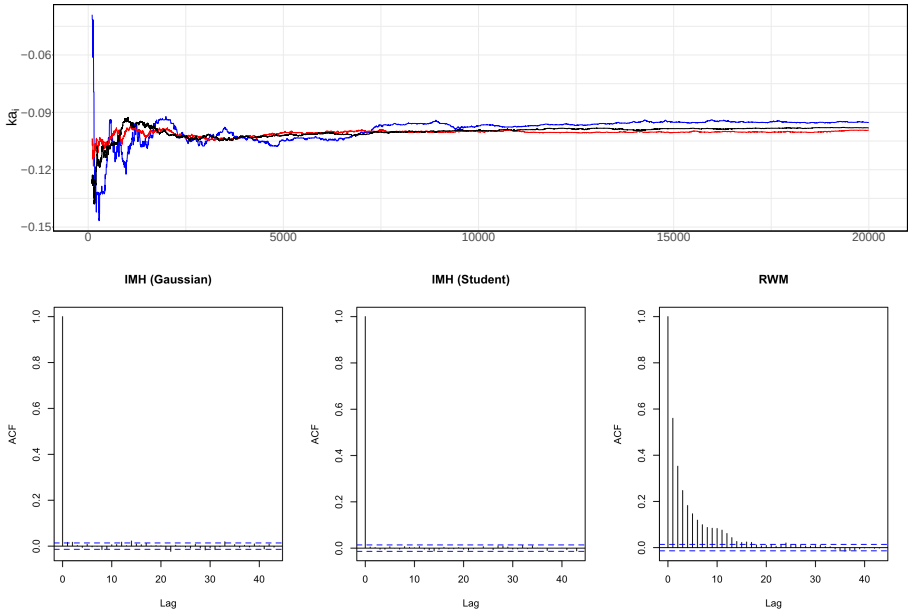

Figure 6.11 – Modelling of the warfarin PK data. Top plot: convergence of the empirical medians of $p_i(ka_i|y_i;\boldsymbol{\theta})$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter ka_i.

Table 6.6 – Pairwise correlations of the proposals.

	ka_i, V_i	ka_i, k_i	V_i, k_i
Variational proposal	0.48	-0.28	-0.61
Laplace proposal	0.56	-0.39	-0.68
NUTS (ground truth)	0.55	-0.39	-0.68

6.8.2 Time-to-event Data Model

Median convergence and autocorrelation plots of the RWM and our nlme-IMH methods for parameter β_i are presented in Figure 6.13. Same observations as for parameter λ_i can

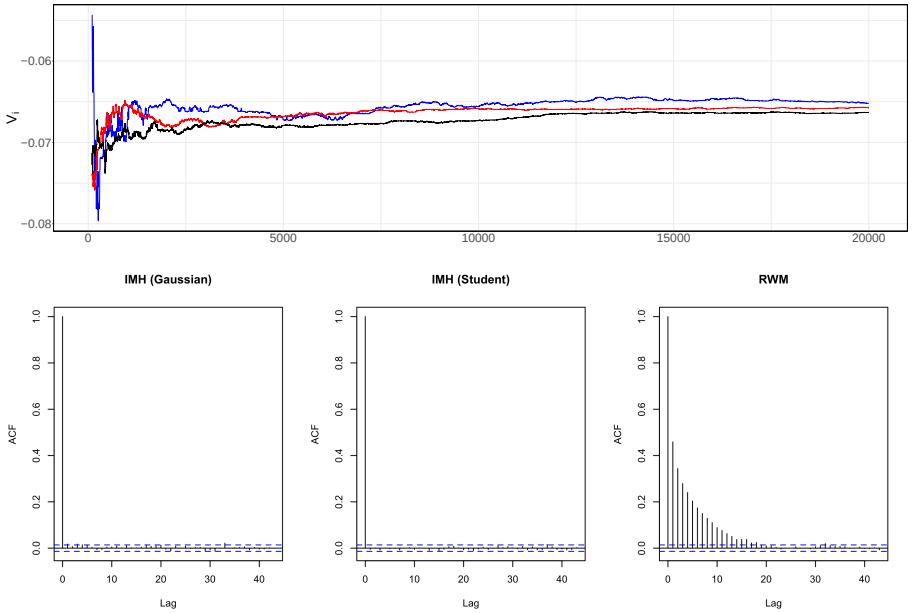

Figure 6.12 – Modelling of the warfarin PK data. Top plot: convergence of the empirical medians of $p_i(V_i|y_i;\boldsymbol{\theta})$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter V_i.

be made.

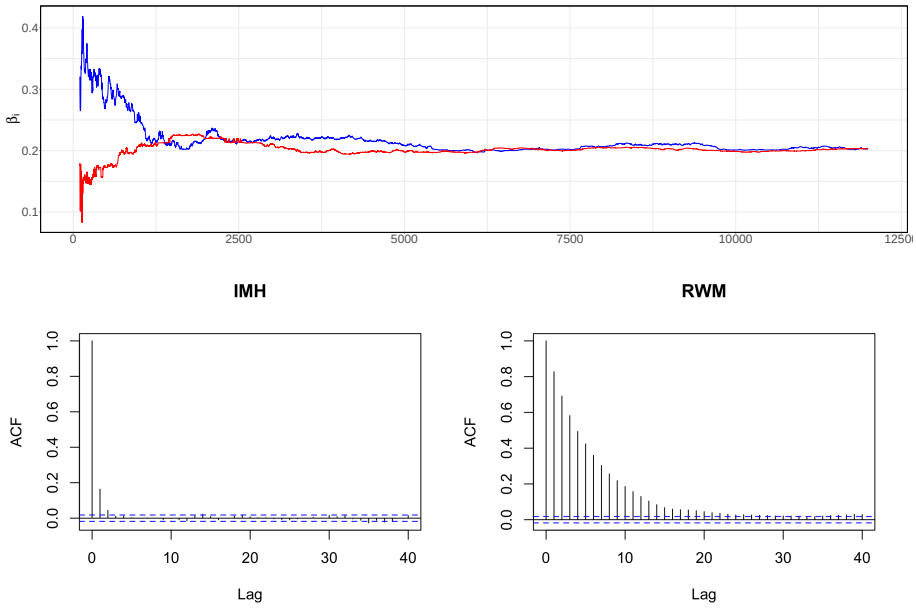

Figure 6.13 – Time-to-event data modelling. Top plot: convergence of the empirical medians of $p_i(\beta_i|y_i;\boldsymbol{\theta})$ for a single individual. Comparison between the reference MH algorithm (blue) and the nlme-IMH (red). Bottom plot: Autocorrelation plots of the MCMC samplers for parameter β_i.

Chapter 7

Incremental Stochastic Approximation of the EM

Abstract: *We develop in this chapter an incremental variant of the SAEM algorithm introduced in Chapter 6. We provide almost sure convergence guaranty of the incremental algorithm and motivate its use through several numerical applications on pharmacokinetics models.*

Contents

7.1 Introduction

We consider a complete model (y, z) where the realisations of y are observed and z is the latent data. When the complete model $f(z, y, \boldsymbol{\theta})$ is parametric, the goal is to compute the maximum likelihood (ML) estimate of the parameter of the incomplete likelihood:

$$\hat{\boldsymbol{\theta}}_{ML} = \arg\max_{\boldsymbol{\theta}\in\Theta} g(y, \boldsymbol{\theta}) \tag{7.1.1}$$

where $\Theta \subset \mathbb{R}^d$ and the incomplete likelihood is defined as:

$$g(y, \boldsymbol{\theta}) = \int_{\mathsf{Z}} f(z, y, \boldsymbol{\theta}) \mathrm{d}z \tag{7.1.2}$$

When the direct derivation of this expression is hard, several methods use the complete model to iteratively find the quantity of interest. The EM algorithm has been the object of considerable interest since its presentation by Dempster, Laird and Rubin in 1977, see [Dempster et al., 1977]. It has been relatively effective in context of maximum likelihood estimation of parameters of incomplete model (unobserved or more). This algorithm is monotonic in likelihood making it a stable tool to work with. This two steps algorithm consists in maximizing an auxiliary quantity that is the expectation of the complete log-likelihood with respect to the conditional distribution over the missing variable conditioned on the observed data and the current parameter estimate (also called the posterior distribution), see [WU, 1983] for more details. Yet, when the quantity computed at the E-step involves unfeasible computations, new methods have been developed in order to by-pass the issue. Most of them alleviate the computation of the expectation using approximates. The Monte Carlo EM (MCEM) algorithm, first introduced in [Wei and Tanner, 1990b], approximates this quantity by a Monte Carlo integration. A Robbins Monroe type approximation can be used to evaluate that latter quantity after the simulation step, that is the SAEM algorithm described in [Lavielle, 1995]. When the posterior distribution of the individual parameters given the observed data is not tractable, sampling from this latter is impossible. The SAEM algorithm is thus coupled with an MCMC procedure to sample latent data from the posterior distribution. Convergence of such an algorithm has been proven in [Kuhn and Lavielle, 2004]. In this article, we study an incremental version of this algorithm where, at each iteration, only a mini-batch of observations are drawn uniformly and considered for updating the optimised quantity.

Two main parts are presented in this Chapter. The first one present the theoretical convergence theorem of the incremental SAEM and the second one highlights the performance of this variant on several numerical examples.

7.2 Maximum Likelihood Estimation: the SAEM Algorithm

7.2.1 Model assumptions and notations

H7.1 *The parameter set Θ is an open subset of $\mathbb{R}^p$.*

We use in the sequel the notations defined in the introductory Chapter 1. Let n be an integer and for $i \in [\![1, n]\!]$, Z be a subset of $\mathbb{R}^m$, μ_i be a σ-finite measure on the

Borel σ-algebra $\mathcal{Z} = \mathcal{B}(\mathsf{Z})$ and $\{f_i(z_i, y_i, \boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$ be a family of positive μ_i-integrable Borel functions on Z. Set $z = (z_i \in \mathsf{Z}, 1 \leq i \leq n)$ and μ the product of the measures $(\mu_i, 1 \leq i \leq n)$. Define, for all $i \in [\![1, n]\!]$ and $\boldsymbol{\theta} \in \Theta$:

$$g_i(y_i, \boldsymbol{\theta}) \triangleq \int_{\mathsf{Z}} f_i(z_i, y_i, \boldsymbol{\theta}) \mu_i(\mathrm{d}z_i) \quad \text{and} \quad p_i(z_i|y_i, \boldsymbol{\theta}) \triangleq \begin{cases} \frac{f_i(z_i, y_i, \boldsymbol{\theta})}{g_i(y_i, \boldsymbol{\theta})} & \text{if } g_i(y_i, \boldsymbol{\theta}) \neq 0 \\ 0 & \text{otherwise} \end{cases} \tag{7.2.1}$$

Note that $p_i(z_i|y_i, \boldsymbol{\theta})$ defines a probability density function with respect to μ_i. Thus $\mathcal{P}_i = \{p_i(z_i|y_i, \boldsymbol{\theta}); \boldsymbol{\theta} \in \Theta\}$ is a family of probability densities. We denote by $\{\mathbb{P}_{i,\boldsymbol{\theta}}; \boldsymbol{\theta} \in \Theta\}$ the associated family of probability measures. For all $\boldsymbol{\theta} \in \Theta$, we set

$$f(z, y, \boldsymbol{\theta}) = \prod_{i=1}^{n} f_i(z_i, y_i, \boldsymbol{\theta}) \quad , \quad g(y, \boldsymbol{\theta}) = \prod_{i=1}^{n} g_i(y_i, \boldsymbol{\theta}) \quad \text{and} \quad p(z|y, \boldsymbol{\theta}) = \prod_{i=1}^{n} p_i(z_i|y_i, \boldsymbol{\theta}) \tag{7.2.2}$$

Our objective is to maximize the function $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta})$ defined as:

$$\mathcal{L}(\boldsymbol{\theta}) \triangleq \log g(y, \boldsymbol{\theta}) = \sum_{i=1}^{n} \mathcal{L}_i(\boldsymbol{\theta}) \tag{7.2.3}$$

where $\mathcal{L}_i(\boldsymbol{\theta}) = \log g_i(y_i, \boldsymbol{\theta})$. The SAEM algorithm is an iterative optimisation algorithm that maximizes the function $\boldsymbol{\theta} \to \mathcal{L}(\boldsymbol{\theta})$ when its direct maximisation is difficult. Define for all $(\boldsymbol{\theta}, \vartheta) \in \Theta^2$:

$$Q_k(\boldsymbol{\theta}, \vartheta) \triangleq \int_{\mathsf{Z}} \log f(z, y, \boldsymbol{\theta}) p(z|y, \vartheta) \mu(\mathrm{d}z) \tag{7.2.4}$$

Denote by $\boldsymbol{\theta}^{(k-1)}$ the current fit of the parameter at iteration k. The k-th step of the SAEM algorithm might be decomposed into three steps:

1. Sampling latent data, for $i \in [\![1, n]\!]$, $z_{i,m}^{(k)} \sim p_i(z_i|y_i; \boldsymbol{\theta}^{(k-1)})$ for $m \in [\![0, M_k - 1]\!]$ under the current model parameter estimate $\boldsymbol{\theta}^{(k-1)}$.

2. Updating the stochastic approximation $\hat{Q}_k(\boldsymbol{\theta})$ of the quantity $Q_k(\boldsymbol{\theta}, \boldsymbol{\theta}^{(k-1)})$, defined by (7.2.4), as follows:

$$\hat{Q}_k(\boldsymbol{\theta}) = \hat{Q}_{k-1}(\boldsymbol{\theta}) + \gamma_k \left[\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \sum_{i=1}^{n} \log f_i(z_{i,m}^{(k)}, y_i; \boldsymbol{\theta}) - \hat{Q}_{k-1}(\boldsymbol{\theta}) \right] \tag{7.2.5}$$

   Where $\{\gamma_k\}_{k>0}$ is a sequence of positive stepsizes.

3. Updating the parameter fit:

$$\boldsymbol{\theta}^{(k)} = \arg\max_{\boldsymbol{\theta} \in \Theta} \hat{Q}_k(\boldsymbol{\theta}) \tag{7.2.6}$$

The SAEM algorithm has been shown theoretically to converge to a maximum of the

likelihood of the observations under very general conditions [Delyon et al., 1999a]. In the simulation step, since the relation between the observed data and the individual parameters can be non linear, sampling from the posterior distribution requires using an inference algorithm. Kuhn et al. in [Kuhn and Lavielle, 2004] proved almost sure convergence of the sequence of parameters obtained by this algorithm coupled with an MCMC procedure during the simulation step. Indeed, $\{z_{i,m}^{(k)}\}_{m=0}^{M_k-1}$ is a Monte Carlo batch. In simple scenarios, the samples $\{z_{i,m}^{(k)}\}_{m=0}^{M_k-1}$ are conditionally independent and identically distributed with distribution $p_i(z_i|y_i,\boldsymbol{\theta}^{(k-1)})$. Nevertheless, in most cases, sampling exactly from this distribution is not an option and the Monte Carlo batch is sampled by Monte Carlo Markov Chains (MCMC) algorithm. MCMC algorithms are a class of methods allowing to sample from complex distribution over (possibly) large dimensional space. In the stochastic approximation step, the sequence of decreasing positive integers γ_k controls the convergence of the algorithm. In practice, γ_k is set equal to 1 during the first K1 iterations to let the algorithm explore the parameter space without memory and to converge quickly to a neighbourhood of the ML estimate. The stochastic approximation is performed during the final K2 iterations where $\gamma_k = 1/k$, ensuring the almost sure convergence of the estimate.

In the sequel, we assume that:

H7.2 *For all $i \in [\![1,n]\!]$, the function $\boldsymbol{\theta} \to g_i(y_i,\boldsymbol{\theta})$ is continuously differentiable on $\boldsymbol{\theta}$ and for all $\boldsymbol{\theta} \in \Theta$:*

$$\nabla_{\boldsymbol{\theta}} g_i(y_i,\boldsymbol{\theta}) = \int_{\mathsf{Z}} \nabla_{\boldsymbol{\theta}} f_i(z_i,y_i,\boldsymbol{\theta}) \mu_i(\mathrm{d}z_i) \tag{7.2.7}$$

and that the model belongs to the curved exponential family:

H7.3 *For all $i \in [\![1,n]\!]$ and $\boldsymbol{\theta} \in \Theta$, The function $f_i(z_i,y_i,\boldsymbol{\theta})$ belongs to the curved exponential family and is given by:*

$$\log f_i(z_i,y_i,\boldsymbol{\theta}) = -\psi_i(\boldsymbol{\theta}) + \langle \tilde{S}_i(z_i,y_i), \phi_i(\boldsymbol{\theta}) \rangle. \tag{7.2.8}$$

where $\psi_i : \boldsymbol{\theta} \mapsto \mathbb{R}$ and $\phi_i : \boldsymbol{\theta} \mapsto \mathbb{R}$ are twice continuously differentiable functions of $\boldsymbol{\theta}$ and $\tilde{S} : \mathsf{Z} \mapsto \mathsf{S}$ is a statistic taking its values in a convex subset S of $\mathbb{R}$ and such that $\int_{\mathsf{Z}} |\tilde{S}_i(z_i,y_i)| p_i(z_i|y_i,\boldsymbol{\theta}) \mu_i(\mathrm{d}z_i) < \infty$.

H7.4 *For all $i \in [\![1,n]\!]$ and $\boldsymbol{\theta} \in \Theta$, the the function $\bar{\mathbf{s}}_i : \boldsymbol{\theta} \to \mathsf{S}_i$ defined as:*

$$\bar{\mathbf{s}}_i(\boldsymbol{\theta}) \triangleq \int_{\mathsf{Z}} \tilde{S}_i(z_i,y_i) p_i(z_i|y_i,\boldsymbol{\theta}) \mu_i(\mathrm{d}z_i) \tag{7.2.9}$$

is continuously differentiable on $\boldsymbol{\theta}$

Define, for all $\boldsymbol{\theta} \in \Theta$ and $s = (s_i, 1 \le i \le n) \in \mathsf{S}$ where $\mathsf{S} = \bigtimes_{n=1}^{N} \mathsf{S}_i$, the function $L(s;\boldsymbol{\theta})$ by:

$$L(s;\boldsymbol{\theta}) \triangleq -\sum_{i=1}^{n} \psi_i(\boldsymbol{\theta}) + \sum_{i=1}^{n} \langle s_i, \phi_i(\boldsymbol{\theta}) \rangle = -\psi(\boldsymbol{\theta}) + \langle s, \phi(\boldsymbol{\theta}) \rangle \tag{7.2.10}$$

where $\psi(\boldsymbol{\theta}) \triangleq \sum_{i=1}^{n} \psi_i(\boldsymbol{\theta})$ and $\phi(\boldsymbol{\theta}) \triangleq (\phi_i(\boldsymbol{\theta}), 1 \leq i \leq n)$.

H7.5 *There exist a function* $\hat{\boldsymbol{\theta}} : \mathsf{S} \mapsto \boldsymbol{\theta}$ *such that for all* $s \in \mathsf{S}$, *:*

$$L(s; \hat{\boldsymbol{\theta}}(s)) \leq L(s; \boldsymbol{\theta}). \tag{7.2.11}$$

where $\hat{\boldsymbol{\theta}}(s)$ *is continuous differentiable on* S.

In many models of practical interest for all $s \in \mathsf{S}$, $\boldsymbol{\theta} \mapsto L(s, \boldsymbol{\theta})$ has a unique minimum. Define the closed set of stationary points $\mathbf{J}$ of $\mathcal{L}(\boldsymbol{\theta})$ as:

$$\mathbf{J} = \{\boldsymbol{\theta} \in \Theta; \nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = 0\} \tag{7.2.12}$$

We can now express the SAEM algorithm for curved exponential family as:

Algorithm 7.1 SAEM algorithm for a curved exponential family

Initialisation: given an initial parameter estimate $\boldsymbol{\theta}^0$, for all $i \in [\![1, n]\!]$ sample initial values $\{z_{i,m}^{(0)}\}_{m=0}^{M_0-1}$ and compute $s_i^0 = \tilde{S}_i(z_i^0, y_i)$.

Iteration k: given the current estimate $\boldsymbol{\theta}^{(k-1)}$:

1. For $i \in [\![1, n]\!]$, sample a Monte Carlo batch $\{z_{i,m}^{(k)}\}_{m=0}^{M_k-1}$ under the current model parameter estimate.
2. Compute $s_i^{(k)}$ such as:

$$s_i^{(k)} = s_i^{(k-1)} + \gamma_k \left(\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \tilde{S}_i(z_{i,m}^{(k)}, y_i) - s_i^{(k-1)}\right) \tag{7.2.13}$$

3. Set $\boldsymbol{\theta}^{(k)} = \hat{\boldsymbol{\theta}}(s^{(k)})$ where $s^{(k)} = (s_i^{(k)}, 1 \leq i \leq n)$.

As mentioned above, convergence properties of this algorithm have been developed in [Delyon et al., 1999a]. They highlight the convergence of the SAEM algorithm depending on the choice of step sizes γ_k and the specification of $M_{(k)}$ used in the stochastic approximation. It is inappropriate to start with small values for step size γ_k and large values for the number of simulations $M_{(k)}$. Rather, it is recommended that one decrease γ_k and increase $M_{(k)}$ as the current approximation of the parameter vector moves closer to a stationary point.

7.2.2 Convergence of the iSAEM for curved exponential family

To avoid cumbersome notations, we set $M_{(k)} = 1$. The incremental version of the SAEM for the curved exponential family can be expressed as:

Algorithm 7.2 Incremental SAEM for a curved exponential family.

Input: given an initial parameter estimate $\boldsymbol{\theta}^0$, for all $i \in [\![1, n]\!]$ sample initial values $\{z_i^{(0)}\}$ and compute $s_i^0 = \tilde{S}_i(z_i^0, y_i)$.
for $k = 0, 1, 2, \ldots, K$ **do**
  Pick a set I_k uniformly on $\{A \subset [\![1, n]\!], \text{card}(A) = p\}$
  For $i \in I_k$, sample $\{z_i^{(k)}\}$ under the current model parameter estimate.
  For $i \in [\![1, n]\!]$, compute $s_i^{(k)}$ such as:

$$s_i^{(k)} = \begin{cases} s_i^{(k-1)} + \gamma_k(\tilde{S}_i(z_i^{(k)}, y_i) - s_i^{(k-1)}) & \text{if } i \in I_k. \\ s_i^{(k-1)} & \text{otherwise.} \end{cases} \tag{7.2.14}$$

  Set $\boldsymbol{\theta}^{(k)} = \hat{\boldsymbol{\theta}}(s^{(k)})$ where $s^{(k)} = (s_i^{(k)}, 1 \leq i \leq n)$.
end for
Return: $\boldsymbol{\theta}^{(k)}$.

We remark that, for all $i \in [\![1, n]\!]$ and $\boldsymbol{\theta} \in \Theta$:

$$s_i^{(k)} = s_i^{(\tau_i^k)} \tag{7.2.15}$$

where for all $i \in [\![1, n]\!]$, $\tau_{i,0} = 0$ and $k \geq 1$ the indices τ_i^k are defined recursively as follows:

$$\tau_i^k = \begin{cases} k & \text{if } i \in I_k \\ \tau_i^{k-1} & \text{otherwise} \end{cases} \tag{7.2.16}$$

Define for all $k \geq 1$:

$$\begin{aligned} h(s^{(k-1)}) &\triangleq \frac{p}{n}\left(\mathbb{E}\left[\tilde{S}(z^{(k)}, y) \,\middle|\, \mathcal{F}_{k-1}\right] - s^{(k-1)}\right) \\ E^{(k)} &\triangleq \bar{I}_k \odot \left(\tilde{S}(z^{(k)}, y) - s^{(k-1)}\right) - \frac{p}{n}\left(\mathbb{E}\left[\tilde{S}(z^{(k)}, y) \,\middle|\, \mathcal{F}_{k-1}\right] - s^{(k-1)}\right) \end{aligned} \tag{7.2.17}$$

where $\odot$ is the Hadamard product, $\bar{I}_k \triangleq (u_i^k, 1 \leq i \leq n)$ and the coefficients u_i^k are defined as follows:

$$u_i^k \triangleq \begin{cases} 1 & \text{if } i \in I_k \\ 0 & \text{otherwise} \end{cases} \tag{7.2.18}$$

and $\mathcal{F}_{k-1}$ is the filtration induced by the sampling of indices and the simulation step up to iteration $k-1$, $\tilde{S}(z^{(k)}, y) := (\tilde{S}_i(z_i^{(k)}, y_i), 1 \leq i \leq n) \in \mathsf{S}$ is the vector of statistics and $s^{(k-1)} = (s_i^{(k-1)}, 1 \leq i \leq n) \in \mathsf{S}$. Since, at iteration k, the iSAEM update can be derived as a Robbins-Monro type update:

$$s^{(k)} = s^{(k-1)} + \gamma_k h(s^{(k-1)}) + \gamma_k E^{(k)} \tag{7.2.19}$$

we recall some convergence properties of a wider class of Robbin-Monro procedure taking the form of:

$$s^{(k)} = s^{(k-1)} + \gamma_k h(s^{(k-1)}) + \gamma_k e^{(k)} + \gamma_k r^{(k)} \tag{7.2.20}$$

where $\{e^{(k)}\}_{k\geq 1}$, the stochastic excitation, and $\{r^{(k)}\}_{k\geq 1}$, the remainder, are random processes defined on the same probability space taking their values in an open subset $\mathcal{H} \subset \mathbb{R}^m$ and h is referring to the mean field of the algorithm. Assume that:

SA 1 $\forall n \geq 0, s^{(k)} \in \mathcal{H}$ *w.p.1*

SA 2 *The sequence of stepsizes* $\{\gamma_k\}_{k\geq 0}$ *is a decreasing sequence of positive numbers such that* $\sum_{k=1}^{\infty} \gamma_k = \infty$

SA 3 *The vector field h is continuous on* $\mathcal{H}$ *and there exists a continuously differentiable function* $V : \mathcal{H} \mapsto \mathbb{R}$ *such that:*

- $\forall s \in \mathcal{H}, F(s) = \langle \nabla_s V(s), h(s) \rangle \leq 0$
- $int(V(\mathbf{J})) = \varnothing$ *where* $\mathbf{J} = \{s \in \mathcal{H} : F(s) = 0\}$

SA 4 *The closure of the set* $\{s^{(k)}\}_{k\geq 1}$ *is a compact subset of* S *w.p.1.*

SA 5 *While considering the RM stochastic approximation procedure ,we can write that the sufficient statistics* $s^{(k)}$ *as follows:*

$$s^{(k)} = s^{(k-1)} + \gamma_k h(s^{(k-1)}) + \gamma_k e^{(k)} + \gamma_k r^{(k)} \tag{7.2.21}$$

$\lim_{p\to\infty} \sum_{k=1}^{p} \gamma_k e^{(k)}$ *exists and is finite,* $\lim_{k\to\infty} r^{(k)} = 0$

We now state the main convergence result of such algorithm:

Theorem 7 *[Delyon et al., 1999a] Assume SA 1 - SA 5. Then we the sequence* $\{s^{(k)}\}_{k>0}$ *from (7.2.20) satisfies:*

$$d(\{s^{(k)}\}_{k>0}, \mathbf{J}) = 0$$

In order to deal with the theoretical convergence properties of the incremental version of the SAEM algorithm, we assume:

iSAEM 1 *The sequence of stepsizes* $\{\gamma_k\}$ *is a decreasing sequence of positive numbers such that* $\forall k > 0, 0 \leq \gamma_k \leq 1, \sum_{k=1}^{\infty} \gamma_k = \infty$ *and* $\sum_{k=1}^{\infty} \gamma_k^2 < \infty$.

iSAEM 2 *For all* $i \in [\![1, n]\!]$ *and any positive Borel function* ϕ_i*:*

$$\mathbb{E}[\phi_i(z_i^{(k+1)})|\mathcal{F}_k] = \int_{\mathsf{Z}} \phi_i(z_i) p_i(z_i|y_i, \boldsymbol{\theta}^{(k)}) \mathrm{d}z_i \tag{7.2.22}$$

Where $\{\mathcal{F}_k\}_{k\geq 0}$ *is the increasing family of* σ*-algebra generated by the random variables up to iteration k.*

iSAEM 3 *For all* $\boldsymbol{\theta} \in \Theta$*:*

$$\sup_{i \in [\![1,n]\!]} \int_{\mathcal{Z}} ||\,\tilde{\mathsf{S}}(z_i)||^2 p_i(z_i|y_i, \boldsymbol{\theta}) \mu(\mathrm{d}z_i) < \infty \tag{7.2.23}$$

and $Cov_{\boldsymbol{\theta}}(\tilde{\mathsf{S}}(z))$ *is continuous with respect to* $\boldsymbol{\theta}$*.*

iSAEM 4 *The functions* $\mathcal{L} : \boldsymbol{\theta} \to \mathbb{R}$ *and* $\hat{\boldsymbol{\theta}} : \mathsf{S} \to \boldsymbol{\theta}$ *are* m *times differentiable*

iSAEM 5 *The closure of* $\{s^{(k)}\}_{k \geq 1}$ *is a compact subset of* S

Lemma 15 *Assume* **H 7.1-H 7.5** *and* **iSAEM4**, *then* **SA3** *is satisfied with* $V(s) \triangleq -\mathcal{L}(\hat{\boldsymbol{\theta}}(s))$*. Also,*

$$\{s \in \mathsf{S} : F(s) = 0\} = \{s \in \mathsf{S} : \nabla_s V(s) = 0\} \tag{7.2.24}$$

$$\hat{\boldsymbol{\theta}}(\{s \in \mathsf{S} : F(s) = 0\}) = \{\boldsymbol{\theta}^* \in \Theta : \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^*) = 0\} \tag{7.2.25}$$

With $F(s) \triangleq \langle \nabla_s V(s), h(s) \rangle$

where the mean-field h is defined in (7.2.17).

Proof The proof is postponed to Appendix 7.5

The main convergence result is expressed as follows:

Theorem 8 *Assume* **H 7.1-H 7.5** *and* **iSAEM1-iSAEM5**, *then the sequence of parameters* $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ *given by Algorithm 7.2 satisfies:*

1. $\lim_{k \to \infty} d(\boldsymbol{\theta}^{(k)}, \mathbf{J}) = 0$
2. $\lim_{k \to \infty} d(s^{(k)}, \{s \in \mathsf{S} : \nabla_s V(s) = 0\}) = 0$

Proof The proof is postponed to Appendix 7.6

This theorem shares the same assumptions of Theorem 5 of [Delyon et al., 1999a]. The main difference, here in the case of the incremental version, resides in the definition of the mean field that will be shown to satisfy assumption SA3.

The results obtained in the previous section demonstrate that, under appropriate conditions, the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k \geq 1}$ converges to a connected component of the solution set $\mathbf{J}$. We assume that those connected components are restricted to points. Then, those converging points could be local minima, maxima or saddle points.

7.3 Numerical Applications

7.3.1 Gaussian Mixture Models

We start by illustrating our findings on a simple GMM model as in Chapter 5. Our goal is to fit a GMM model to a set of n observations $\{y_i\}_{i=1}^n$ whose distribution is modeled as a Gaussian mixture of V components, each with a unit variance. Let $z_i \in [\![M]\!]$ be the latent labels, the complete log-likelihood is:

$$\log f(z_i, y_i; \boldsymbol{\theta}) = \sum_{v=1}^V \mathbb{1}_{\{v\}}(z_i) \left[\log(\omega_v) - \mu_v^2/2\right] + \sum_{v=1}^V \mathbb{1}_{\{v\}}(z_i)\mu_v y_i + \text{constant} \ . \quad (7.3.1)$$

where $\boldsymbol{\theta} := (\boldsymbol{\omega}, \boldsymbol{\mu})$ with $\boldsymbol{\omega} = \{\omega_v\}_{v=1}^{V-1}$ are the mixing weights with the convention $\omega_V = 1 - \sum_{v=1}^{V-1} \omega_v$ and $\boldsymbol{\mu} = \{\mu_v\}_{v=1}^V$ are the means. We use the penalization $\mathrm{R}(\boldsymbol{\theta}) = \frac{\delta}{2}\sum_{v=1}^V \mu_v^2 - \log \mathrm{Dir}(\boldsymbol{\omega}; V, \epsilon)$ where $\delta > 0$ and $\mathrm{Dir}(\cdot; V, \epsilon)$ is the V dimensional symmetric Dirichlet distribution with concentration parameter $\epsilon > 0$. The constraint set on $\boldsymbol{\theta}$ is given by

$$\Theta = \{\omega_v, \ v = 1, ..., V-1 : \omega_v \geq 0, \ \textstyle\sum_{v=1}^{V-1} \omega_v \leq 1\} \times \{\mu_v \in \mathbb{R}, \ v = 1, ..., V\}. \quad (7.3.2)$$

Model assumptions Using the partition of the sufficient statistics as $S(y_i, z_i) = (S^{(1)}(y_i, z_i)^\top, S^{(2)}(y_i, z_i)^\top, S^{(3)}(y_i, z_i))^\top \in \mathbb{R}^{V-1} \times \mathbb{R}^{V-1} \times \mathbb{R}$, the partition $\phi(\boldsymbol{\theta}) = (\phi^{(1)}(\boldsymbol{\theta})^\top, \phi^{(2)}(\boldsymbol{\theta})^\top, \phi^{(3)}(\boldsymbol{\theta}))^\top \in \mathbb{R}^{V-1} \times \mathbb{R}^{V-1} \times \mathbb{R}$ and the fact that $\mathbb{1}_{\{V\}}(z_i) = 1 - \sum_{v=1}^{V-1} \mathbb{1}_{\{v\}}(z_i)$, the complete data log-likelihood can be expressed as in (5.1.2) with

$$\begin{aligned}
&s_{i,m}^{(1)} = \mathbb{1}_{\{v\}}(z_i), \quad \phi_v^{(1)}(\boldsymbol{\theta}) = \left\{\log(\omega_v) - \frac{\mu_v^2}{2}\right\} - \left\{\log(1 - \textstyle\sum_{j=1}^{V-1}\omega_j) - \frac{\mu_V^2}{2}\right\} \ , \\
&s_{i,m}^{(2)} = \mathbb{1}_{\{v\}}(z_i) y_i, \quad \phi_v^{(2)}(\boldsymbol{\theta}) = \mu_v \ , \\
&s_i^{(3)} = y_i, \quad \phi^{(3)}(\boldsymbol{\theta}) = \mu_V \ ,
\end{aligned} \quad (7.3.3)$$

and $\psi(\boldsymbol{\theta}) = -\left\{\log(1 - \sum_{v=1}^{V-1} \omega_v) - \frac{\mu_V^2}{2\sigma^2}\right\}$. We also define for each $v \in [\![1, V]\!]$, $j \in [\![1, 3]\!]$, $s_v^{(j)} = n^{-1}\sum_{i=1}^n s_{i,v}^{(j)}$. Consider the following conditional expected value:

$$\widetilde{\omega}_v(y_i; \boldsymbol{\theta}) := \mathbb{E}_{\boldsymbol{\theta}}[\mathbb{1}_{\{z_i = v\}} | y = y_i] = \frac{\omega_v \exp(-\frac{1}{2}(y_i - \mu_i)^2)}{\sum_{j=1}^V \omega_j \ \exp(-\frac{1}{2}(y_i - \mu_j)^2)} \ , \quad (7.3.4)$$

where $v \in [\![1, V]\!]$, $i \in [\![1, n]\!]$ and $\boldsymbol{\theta} = (\boldsymbol{w}, \boldsymbol{\mu}) \in \Theta$. In particular, given $\boldsymbol{\theta} \in \Theta$, the E-step updates of the quantity defined in (7.2.9) can be written as

$$\bar{\mathbf{s}}_i(\boldsymbol{\theta}) = (\underbrace{\widetilde{\omega}_1(y_i; \boldsymbol{\theta}), ..., \widetilde{\omega}_{V-1}(y_i; \boldsymbol{\theta})}_{:= \bar{\mathbf{s}}_i^{(1)}(\boldsymbol{\theta})^\top}, \underbrace{y_i\widetilde{\omega}_1(y_i; \boldsymbol{\theta}), ..., y_i\widetilde{\omega}_V(y_i; \boldsymbol{\theta})}_{:= \bar{\mathbf{s}}_i^{(2)}(\boldsymbol{\theta})^\top}, \underbrace{y_i}_{:= \bar{\mathbf{s}}_i^{(3)}(\boldsymbol{\theta})})^\top. \quad (7.3.5)$$

Recall that we have used the following regularizer:

$$\mathrm{R}(\boldsymbol{\theta}) = \tfrac{\delta}{2}\sum_{v=1}^{V}\mu_v^2 - \epsilon\sum_{v=1}^{V}\log(\omega_v) - \epsilon\log\left(1-\sum_{v=1}^{V-1}\omega_v\right), \tag{7.3.6}$$

It can be shown that the regularized M-step in (5.2.5) evaluates to

$$\overline{\boldsymbol{\theta}}(\boldsymbol{s}) = \begin{pmatrix} (1+\epsilon V)^{-1}\big(s_1^{(1)}+\epsilon,\ldots,s_{V-1}^{(1)}+\epsilon\big)^\top \\ \big((s_1^{(1)}+\delta)^{-1}s_1^{(2)},\ldots,(s_{V-1}^{(1)}+\delta)^{-1}s_{V-1}^{(2)}\big)^\top \\ \big(1-\sum_{v=1}^{V-1}s_v^{(1)}+\delta\big)^{-1}\big(s^{(3)}-\sum_{v=1}^{V-1}s_v^{(2)}\big) \end{pmatrix} = \begin{pmatrix} \overline{\boldsymbol{\omega}}(\boldsymbol{s}) \\ \overline{\boldsymbol{\mu}}(\boldsymbol{s}) \\ \overline{\mu}_V(\boldsymbol{s}) \end{pmatrix}. \tag{7.3.7}$$

where we have defined for all $v \in [\![1,V]\!]$ and $j \in [\![1,3]\!]$, $s_v^{(j)} = n^{-1}\sum_{i=1}^{n} s_{i,v}^{(j)}$.

Algorithms updates In the sequel, for all $i \in [\![n]\!]$ and iteration k, the conditional expectation $\overline{\mathbf{s}}_i^{(k)}$ and is equal to:

$$\overline{\mathbf{s}}_i^{(k)} = \begin{pmatrix} \big(\widetilde{\omega}_1(y_i;\boldsymbol{\theta}^{(k)}),\ldots,\widetilde{\omega}_{V-1}(y_i;\boldsymbol{\theta}^{(k)})\big)^\top \\ \big(y_i\widetilde{\omega}_1(y_i;\boldsymbol{\theta}^{(k)}),\ldots,y_i\widetilde{\omega}_{V-1}(y_i;\boldsymbol{\theta}^{(k)})\big)^\top \\ y_i \end{pmatrix}. \tag{7.3.8}$$

At iteration k, the several E-steps defined by (5.2.4) or (5.2.5) or (5.2.6) or (5.2.7) leads to the definition of the quantity $\hat{\mathbf{s}}^{(k+1)}$. For the GMM example, after the initialization of the quantity $\hat{\mathbf{s}}^{(0)} = n^{-1}\sum_{i=1}^{n}\overline{\mathbf{s}}_i^{(0)}$, those E-steps break down as follows:

Batch EM (EM): for all $i \in [\![1,n]\!]$, compute $\overline{\mathbf{s}}_i^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = n^{-1}\sum_{i=1}^{n}\overline{\mathbf{s}}_i^{(k)}. \tag{7.3.9}$$

Incremental EM (iEM): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \overline{\mathbf{s}}_{i_k}^{(k)} - \overline{\mathbf{s}}_{i_k}^{(\tau_i^k)} = n^{-1}\sum_{i=1}^{n}\overline{\mathbf{s}}_i^{(\tau_i^k)}. \tag{7.3.10}$$

Online EM (sEM): draw an index i_k uniformly at random on $[\![n]\!]$, compute $\overline{\mathbf{s}}_{i_k}^{(k)}$ and set

$$\hat{\mathbf{s}}^{(k+1)} = (1-\gamma_k)\hat{\mathbf{s}}^{(k)} + \gamma_k\overline{\mathbf{s}}_{i_k}^{(k)}. \tag{7.3.11}$$

Incremental SAEM (iSAEM): draw an index i_k uniformly at random on $[\![n]\!]$, draw $z_{i_k}^{(k)}$ from its conditional distribution $p_i(z_i|y_i,\boldsymbol{\theta}^{(k-1)})$ and set

$$\hat{\mathbf{s}}^{(k+1)} = \hat{\mathbf{s}}^{(k)} + \gamma_k\big(\tilde{S}_{i_k}(z_{i_k}^{(k)},y_{i_k}) - \overline{\mathbf{s}}_{i_k}^{(\tau_i^k)}\big). \tag{7.3.12}$$

where $\tilde{S}_{i_k}(z_{i_k}^{(k)},y_{i_k}) = (z_{i_k}^{(k)}, y_{i_k}z_{i_k}^{(k)}, y_{i_k})$. The mini-batch version of the iSAEM boils down

from the last update. Finally, apply the maximization step to yield the new parameter estimate $\boldsymbol{\theta}^{(k+1)} = \overline{\boldsymbol{\theta}}(\hat{\mathbf{s}}^{(k+1)})$.

Experimental results We generate $n = 10^3$ samples from a GMM model with $M = 2$ components with two mixtures with means $\mu_1 = -\mu_2 = 0.5$. All plots below are generated averaging over 15 independent simulated datasets. We use $n = 10^4$ synthetic samples and run the EM method until convergence (to double precision) to obtain the ML estimate $\mu^\star$. We compare the EM, iEM , sEM (an online version of the EM developed by Cappé and Moulines [2009] and iSAEM methods in terms of their precision measured by $|\mu - \mu^\star|^2$. We set the stepsize of the iSAEM as $\gamma_k = 1/k^{0.6}$ and average over $M_{(k)} = 30$ MC samples for the iSAEM runs. Figure 7.1 shows the convergence of the precision $|\mu - \mu^*|^2$ for the different methods against the epoch(s) elapsed (one epoch equals n iterations). We observe that the iSAEM and iEM methods outperform the batch methods. Though, the iSAEM algorithm, after a certain number of epochs, seems to have reached its maximum precision. High variance, due to its simulation step, prevents the algorithm to attain higher precision as in the iEM algorithm.

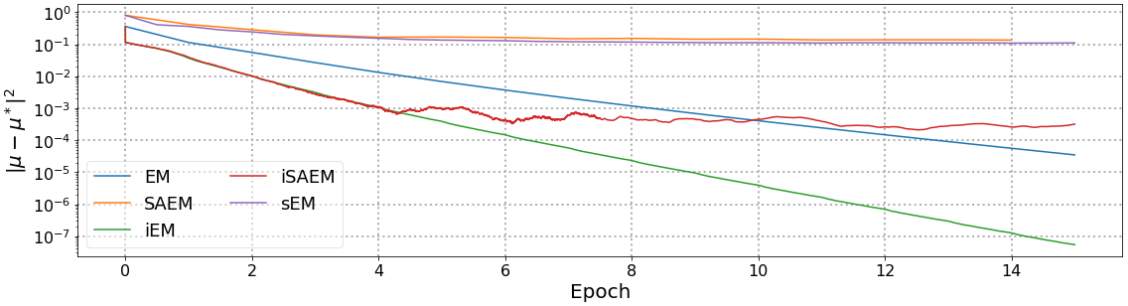

Figure 7.1 – Performance of EM and SAEM methods for fitting a GMM: Precision ($|\mu^{(k)} - \mu^\star|^2$) as a function of the epoch elapsed.

7.3.2 Pharmacokinetic model

We consider the same PK model studied Section 6.5 of Chapter 6 We recall that the model is a one-compartment pharmacokinetics (PK) model for oral administration, assuming first-order absorption and linear elimination processes:

$$f(t, ka, V, k) = \frac{D\, ka}{V(ka - k)}(\mathrm{e}^{-ka\, t} - \mathrm{e}^{-k\, t}) \,, \tag{7.3.13}$$

where ka is the absorption rate constant, V the volume of distribution , k the elimination rate constant, and D the dose of drug administered. Here, ka, V and k are PK parameters that can change from one individual to another. We note $\psi_i = (ka_i, V_i, k_i)$ be the vector of individual PK parameters for individual i. We assume in this example that the residual errors are independent and normally distributed with mean 0 and variance σ^2 and that

lognormal distributions are used for the three PK parameters:

$$\log(ka_i) \sim \mathcal{N}(\log(ka_{\text{pop}}), \omega_{ka}^2) \ , \log(V_i) \sim \mathcal{N}(\log(V_{\text{pop}}), \omega_V^2) \ , \log(k_i) \sim \mathcal{N}(\log(k_{\text{pop}}), \omega_k^2) \ .$$

To illustrate the behaviors of the iSAEM algorithm, we generate a dataset of $n = 10^3$ individuals with $n_i = 5$ observations per individual using the PK model described above. We use the following generating parameters:$ka_{\text{pop}} = 2$, $V_{\text{pop}} = 10$, $k_{\text{pop}} = 1$, $\omega_{ka} = 0.3$, $\omega_V = 0.2$, $\omega_k = 0.1$ and $\sigma^2 = 1$. We then run the SAEM and the iSAEM algorithms for 200 iterations using a stepsize $\gamma_k = 1$ and then 50 iterations using $\gamma_k = 1/k$ to ensure almost sure convergence.

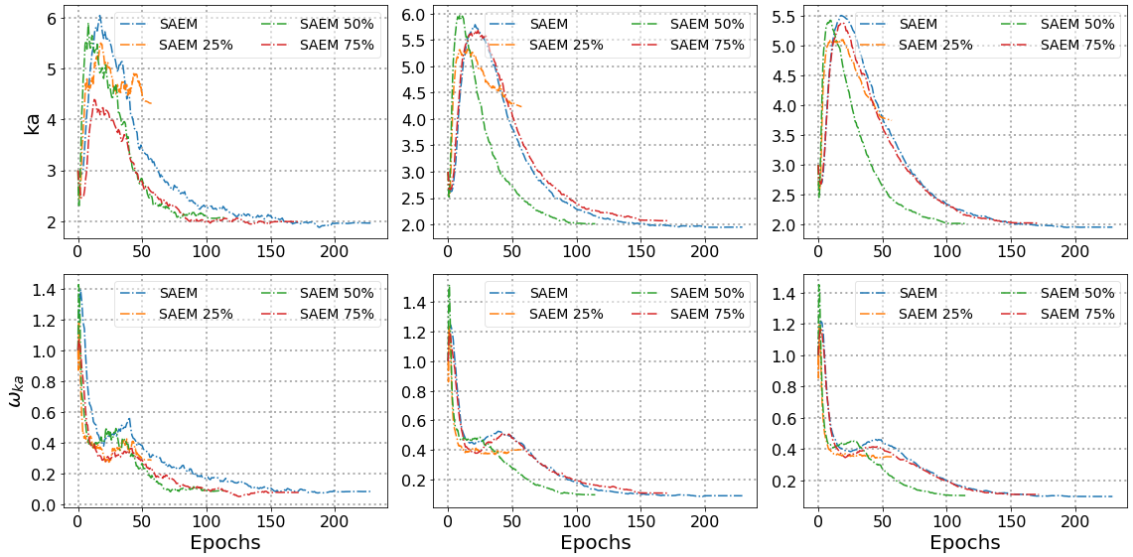

Figure 7.2 – Estimation of the population PK parameters: convergence of the sequences of estimates $(\hat{ka}_{\text{pop}}^{(k)}, 1 \leq k \leq 250)$ and $(\hat{\omega}_{ka}^{(k)}, 1 \leq k \leq 250)$ obtained with the SAEM and the iSAEM algorithms. From left to right, the runs are executed averaging over, respectively, 1, 10 and 20 chains.

Figure 7.2 shows the effect of the Monte Carlo batch size on the convergence behaviors of the incremental variants of the SAEM. While no distinct differences can be observed when the updates use only one sample from the conditional distribution, one can notice that increasing the number of samples not only smooth the convergence of the estimated parameters but also exhibits a different rates according to the number of individuals picked at each iteration. These empirical results illustrate a tradeoff between resorting to cheap updates, independent of the problem size in general, in order to reduce the bias after a small number of iterations, and decreasing the variance of the estimates, that generally explodes with small mini-batch size.

An attempt to illustrate the aforementioned bias-variance tradeoff on a simple linear Gaussian example can be found in Appendix 7.7.

7.4 Conclusion

We have presented in this chapter an incremental variant of the SAEM algorithm. Based upon important convergence results for stochastic approximation scheme, we establish the almost sure convergence of our incremental scheme. We illustrate our findings through numerical experiments on a simple GMM case and a more sophisticated pharmacokinetics example. Not only do we observe asymptotic convergence, as proven in this Chapter, but we also exhibit a bias-variance tradeoff due to the interplay between the variance caused by the Monte Carlo batch and the mini-batch of picked individuals at each iteration. Non-asymptotic bounds for this algorithm would be the next important result to derive along with a thorough study of the influence of the mini-batch size on the convergence rate.

Appendices to Incremental Stochastic Approximation of the EM

7.5 Proof of Lemma 15

Lemma *Assume* **H 7.1-H 7.5** *and* **iSAEM4**, *then* **SA3** *is satisfied with* $V(s) \triangleq -\mathcal{L}(\hat{\boldsymbol{\theta}}(s))$. *Also,*

$$\{s \in \mathsf{S} : F(s) = 0\} = \{s \in \mathsf{S} : \nabla_s V(s) = 0\} \tag{7.5.1}$$

$$\hat{\boldsymbol{\theta}}(\{s \in \mathsf{S} : F(s) = 0\}) = \{\boldsymbol{\theta}^* \in \Theta : \nabla_{\boldsymbol{\theta}} l(\boldsymbol{\theta}^*) = 0\} \tag{7.5.2}$$

With $F(s) \triangleq \langle \nabla_s V(s), h(s) \rangle$

Proof In this proof, one need to express the mean field of the iSAEM algorithm at each iteration and propose a Lyapunov function that satisfies SA 3. It is well known (see [Delyon et al., 1999a]) that the incomplete data likelihood is a Lyapunov function relative to the SAEM mapping. We will show that this same function is a Lyapunov function relative to the incremental SAEM mapping. The k-th iteration of the iSAEM algorithm is expressed as:

$$s^{(k)} = s^{(k-1)} + \gamma_k \bar{I}_k \odot \left(\tilde{S}(z^{(k)}, y) - s^{(k-1)} \right) \tag{7.5.3}$$

Given that:

$$\begin{aligned} \mathbb{E}\left[\bar{I}_k \odot \tilde{S}(z^{(k)}, y) \,\middle|\, \hat{\boldsymbol{\theta}}(s^{(k-1)}) \right] &= \frac{p}{n} \mathbb{E}\left[\tilde{S}(z^{(k)}, y) \,\middle|\, \hat{\boldsymbol{\theta}}(s^{(k-1)}) \right] \\ \mathbb{E}\left[\bar{I}_k \odot s^{(k-1)} \,\middle|\, \hat{\boldsymbol{\theta}}(s^{(k-1)}) \right] &= \frac{p}{n} s^{(k-1)} \end{aligned} \tag{7.5.4}$$

we can express the mean field $h(s)$ of the algorithm defined for all $s \in \mathsf{S}$ as:

$$h(s) = \frac{p}{n} \left(\mathbb{E}\left[\tilde{S}(z) \,\middle|\, \hat{\boldsymbol{\theta}}(s) \right] - s^{(k-1)} \right) = \frac{p}{n} \left(\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}(s)) - s \right) \tag{7.5.5}$$

Under M7.4 and iSAEM 4 we obtain that $h(s)$ is continuously differentiable on S. Under,

M7.5, $\hat{\boldsymbol{\theta}}(s)$ is a solution of the maximization of $L(s, \boldsymbol{\theta})$, thus:

$$\begin{aligned} &\nabla_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) = 0 \\ \Rightarrow &- \nabla_{\boldsymbol{\theta}} \psi(\hat{\boldsymbol{\theta}}(s)) + s^t \nabla_{\boldsymbol{\theta}} \phi(\hat{\boldsymbol{\theta}}(s)) = 0 \end{aligned} \tag{7.5.6}$$

The differentiation of the latter with respect to the vector s (under assumptions M7.3 and M7.5) yields:

$$\begin{aligned} &\nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) \nabla_s \hat{\boldsymbol{\theta}}(s) + \nabla_{\boldsymbol{\theta}} \phi(\hat{\boldsymbol{\theta}}(s))^t = 0 \\ \Rightarrow &\nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) \nabla_s \hat{\boldsymbol{\theta}}(s) = -\nabla_{\boldsymbol{\theta}} \phi(\hat{\boldsymbol{\theta}}(s))^t \end{aligned} \tag{7.5.7}$$

Also , under M7.2, the Fisher identity [Fisher, 1925] gives:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = \int \nabla_{\boldsymbol{\theta}} \log f(z, y, \boldsymbol{\theta}) p(z|y, \boldsymbol{\theta}) \mu(\mathrm{d}z) \tag{7.5.8}$$

which rewrites:

$$\nabla_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta}) = -\nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta}) + \bar{\mathbf{s}}(\boldsymbol{\theta})^t \nabla_{\boldsymbol{\theta}} \phi(\boldsymbol{\theta}) \tag{7.5.9}$$

Define the following Lyapunov function and recall the mean field $h(s)$ for all $s \in \mathsf{S}$:

$$V(s) = -\mathcal{L}(\hat{\boldsymbol{\theta}}(s)) \quad \text{and} \quad h(s) = \frac{p}{n} \left(\bar{\mathbf{s}}(\hat{\boldsymbol{\theta}}(s)) - s \right) \tag{7.5.10}$$

We are going to show that for all $s \in \mathsf{S}$, $F(s) \triangleq \langle \nabla_s V(s), h(s) \rangle < 0$. Plugging (7.5.6) into (7.5.9) gives:

$$\begin{aligned} \frac{p}{n} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}(s)) &= \underbrace{(\frac{p}{n}(\bar{\mathbf{s}} - s))^t}_{h(s)} \underbrace{\nabla_{\boldsymbol{\theta}} \phi(\hat{\boldsymbol{\theta}}(s))}_{-\nabla_s \hat{\boldsymbol{\theta}}(s)^t \nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s))} \\ &= -h(s)^t \nabla_s \hat{\boldsymbol{\theta}}(s)^t \nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) \end{aligned} \tag{7.5.11}$$

We can derive this expression with respect to the vector s. The gradient of $\mathcal{L}(\hat{\boldsymbol{\theta}}(s))$ is given by the following relation:

$$\begin{aligned} \frac{p}{n} \nabla_s \mathcal{L}(\hat{\boldsymbol{\theta}}(s)) &= \frac{p}{n} \nabla_{\boldsymbol{\theta}} \mathcal{L}(\hat{\boldsymbol{\theta}}(s)) \nabla_s \hat{\boldsymbol{\theta}}(s)) \\ &= -h(s)^t \nabla_s \hat{\boldsymbol{\theta}}(s)^t \nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) \nabla_s \hat{\boldsymbol{\theta}}(s) \end{aligned} \tag{7.5.12}$$

The quantity of interest can be expressed as:

$$\begin{aligned} F(s) = \langle \nabla_s V(s), h(s) \rangle &= -\langle \nabla_s \mathcal{L}(\hat{\boldsymbol{\theta}}(s)), h(s) \rangle \\ &= \frac{n}{p} h(s)^t \nabla_s \hat{\boldsymbol{\theta}}(s)^t \nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) \nabla_s \hat{\boldsymbol{\theta}}(s) h(s) \end{aligned} \tag{7.5.13}$$

Since, under assumption M7.5, $\nabla^2_{\boldsymbol{\theta}} L(s, \hat{\boldsymbol{\theta}}(s)) \leq 0$ we have that $\langle \nabla_s V(s), h(s) \rangle \leq 0$ which proves the first part of SA 3.

Obviously, $\{s \in \mathsf{S} : \nabla_s V(s) = 0\} \subset \{s \in \mathsf{S} : F(s) = 0\}$.
If $s^* \in \{s \in \mathsf{S} : F(s) = 0\}$ then:

$$\begin{aligned} F(s^*) &= \frac{n}{p} h(s)^t \nabla_s \hat{\boldsymbol{\theta}}(s^*)^t \nabla^2_{\boldsymbol{\theta}} L(s^*, \hat{\boldsymbol{\theta}}(s^*)) \nabla_{s^*} \hat{\boldsymbol{\theta}}(s^*) h(s^*) \\ &= \langle \nabla_s \mathcal{L}(\hat{\boldsymbol{\theta}}(s)), h(s) \rangle = 0 \end{aligned} \tag{7.5.14}$$

Since $\nabla^2_{\boldsymbol{\theta}} L(s^*, \hat{\boldsymbol{\theta}}(s^*))$ is non positive then $\nabla_s \mathcal{L}(\hat{\boldsymbol{\theta}}(s^*)) = 0$ which proves $\nabla_s V(s^*) = 0$ and the reverse inclusion. Using Sard's theorem, in [Brocker et al., 1975], we have that $V(\{s \in \mathsf{S}, \nabla_s V(s) = 0\})$ has zero Lebesgue measure which proves the second part of 15. ■

7.6 Proof of Theorem 8

Theorem *Assume* **H 7.1-H 7.5** *and* **iSAEM 1-iSAEM 5**, *then the sequence of parameters* $\{\boldsymbol{\theta}^{(k)}\}_{k>0}$ *given by Algorithm 7.2 satisfies:*

1. $\lim_{k\to\infty} d(\boldsymbol{\theta}^{(k)}, \mathbf{J}) = 0$
2. $\lim_{k\to\infty} d(s^{(k)}, \{s \in \mathsf{S} : \nabla_s V(s) = 0\}) = 0$

Proof First of all, we verify assumptions of Theorem 2. SA1 is verified under M7.1 and iSAEM 1 because the stepsize γ_k is strictly inferiror to 1 and the convex hull of $\tilde{\mathsf{S}}(\mathbb{R}^p)$ is in S. SA2 is implied by iSAEM 1 and SA4 by iSAEM 5. Note that under iSAEM 5, there exists w.p.1 a compact set $\mathcal{K}$, such that $s_k \in \mathcal{K}$ for all $k \geq 0$. Denote $M^{(n)} = \sum_{k=1}^n \gamma_k E^{(k)}$. Then $\{M^{(n)}\}_{n\geq 1}$ is a martingale which satisfies, under iSAEM 1, iSAEM 2 and iSAEM 4:

$$\begin{aligned} &\sum_{n=1}^{\infty} \mathbb{E}\Big[||M^{(n+1)} - M^{(n)}||^2 | \mathcal{F}_n\Big] \\ &= \sum_{n=1}^{\infty} \mathbb{E}\Big[||\gamma_{n+1} E^{(n+1)}||^2 | \mathcal{F}_n\Big] \\ &\leq \frac{p}{n} \sum_{n=1}^{\infty} \gamma_{n+1}^2 \sum_{i=1}^{n} \left\{ \int ||\tilde{S}_i(z_i)||^2 p_i(z_i, \hat{\boldsymbol{\theta}}(s_{i_n+1})) \mu_i(\mathrm{d}z_i) \right\} < \infty \end{aligned} \tag{7.6.1}$$

This proves the existence of $\lim M^{(n)}$, see [Hall and Heyde, 2014, Theorem 2.15 p.33]. Assumption SA3 is thus verified using 15. We check the assumptions of Theorem 7 which yields:

$$d(\boldsymbol{\theta}^{(k)}, \mathbf{J}) \to 0 \quad \text{w.p.1} \tag{7.6.2}$$

The second part of the Theorem is proved by applying Lemma 15. ■

7.7 Bias-Variance Tradeoff in Incremental EM and SAEM

The recent development of incremental techniques involves faster gradient descent algorithms. The original full gradient descent combined with the stochastic version to propose an averaged gradient solution [Defazio et al., 2014, Roux et al., 2012] for the strongly convex sum of a finite set of smooth functions. It incorporates a memory of previous gradients at each iteration to reach a faster convergence rate.

Preliminary remarks regarding incremental algorithms: In this section, we will focus on how to practically implement this algorithm. Two main tuning parameters need to be chosen. The first one being the size of the batch of indices considered at each iteration and the second is the strategy of choosing those indices. Indeed, following numerous improvement of the gradient descent algorithm for instance, justifying the incremental choice of data sample at each iteration, considering incremental of individuals in the context of mixed effects models makes sense. Also, usually those incremental version of existing algorithm showcase picking the individuals according to a uniform distribution but accelerated versions introduce different choice strategy depending on a well chosen score parameter. Roux et al. [2012], for instance, consider at each iteration, of their averaged version of the stochastic gradient descent, the index whose gradient is the highest. Of course this strategy is optimal and requires computing all gradients at each iteration which can be costly in the context of high dimensional data.

Optimal batch size for a simple linear Gaussian model: Let us consider the case when all the variables of interest are Gaussian.

$$y_i = z_i + \epsilon_i \ , \tag{7.7.1}$$

where $z_i \sim \mathcal{N}(\boldsymbol{\theta}, \omega^2)$ and $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$. Since the z_i and ϵ_i are i.i.d we have that $y_i \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 + \omega^2)$ and $y_i|z_i \sim \mathcal{N}(z_i, \sigma^2)$. The goal is to find an estimate of the mean $\boldsymbol{\theta}$ that maximizes the likelihood $p(y, \boldsymbol{\theta})$ considering that σ^2 and ϵ^2 are known. The maximum likelihood is easy to compute in this case since $y_i \sim \mathcal{N}(\boldsymbol{\theta}, \sigma^2 + \omega^2)$:

$$\boldsymbol{\theta}_{ML} = \frac{1}{n} \sum_{i=1}^{n} y_i \tag{7.7.2}$$

We can rewrite the complete log likelihood $\log f_i(z, y, \boldsymbol{\theta})$ as part of the exponential family:

$$\begin{aligned} \log f_i(z, y, \boldsymbol{\theta}) &= \sum_{i=1}^{n} \left(\log p_i(y_i|z_i, \boldsymbol{\theta}) + \log p_i(z_i, \boldsymbol{\theta}) \right) \\ &= \sum_{i=1}^{n} -\frac{1}{2} \log(2\pi\sigma^2) - \frac{(y_i - z_i)^2}{2\sigma^2} - \frac{1}{2} \log(2\pi\omega^2) - \frac{(z_i - \boldsymbol{\theta})^2}{2\omega^2} \end{aligned} \tag{7.7.3}$$

The resulting statistics are:

$$S_1(y,z) = \sum_{i=1}^{n} z_i, \quad S_2(y,z) = \sum_{i=1}^{n} z_i y_i \quad \text{and} \quad S_3(y,z) = \sum_{i=1}^{n} z_i^2 \tag{7.7.4}$$

Let us define the quantity of interest $p_i(z_i|y_i,\boldsymbol{\theta})$ using Bayes rule. We find that $z_i|y_i \sim \mathcal{N}(\alpha\boldsymbol{\theta} + (1-\alpha)\bar{y}, \Gamma^2)$ with $\alpha = \frac{\sigma^2}{\sigma^2+\omega^2}$ and $\Gamma^2 = \frac{\sigma^2\omega^2}{\sigma^2+\omega^2}$.

Incremental EM algorithm In the general case, where we consider that we pick a batch size of size pN at each iteration, where $p \in [\![0,1]\!]$, then the general recurrent relation between parameter estimates is:

$$\boldsymbol{\theta}^{(k)} = \rho_p^{1/p}\boldsymbol{\theta}^{(k-1/p)} + (1-\alpha)\bar{y}e_1 \tag{7.7.5}$$

where:

$$\rho_p = \begin{pmatrix} \frac{\alpha p}{n} & .. & .. & \frac{\alpha p}{n} \\ 1 & 0 & .. & 0 \\ 0 & 1 & .. & 0 \\ .. & .. & .. & .. \\ .. & .. & .. & 0 \end{pmatrix} \quad \text{and} \quad \underline{e}_1 = \begin{pmatrix} 1 \\ 0 \\ .. \\ 0 \end{pmatrix} \tag{7.7.6}$$

What is really important here are the eigenvalues of ρ at the power $\frac{1}{p}$. These are the values that will drive the speed of convergence. Besides, the highest eigenvalue is enough to compare the rate of two algorithms (for instance, for two different values of mini-batch size $p1$ and $p2$). We denote $\Lambda(p) = (\lambda_p)^{1/p}$ where $\lambda_p = \max(\text{eigenvalues}(\rho_p))$ and calculate the characteristic polynomial of ρ:

$$P_{\rho_p}(X) = (-1)^{1/p}(x^{1/p} - \frac{\alpha p}{n}\sum_{i=0}^{1/p-1} x^i) \tag{7.7.7}$$

Naturally $P_{\rho_p}(\lambda_p) = 0$ so:

$$P_{\rho_p}(\Lambda(p)^p) = 0 = (-1)^{1/p}(\Lambda(p) - \frac{\alpha p}{n}\sum_{i=0}^{1/p-1} (\Lambda(p)^{1/p})^i) \tag{7.7.8}$$

Since $0 < \Lambda(p) < 1$ we have that :

$$\begin{aligned} &(-1)^{1/p}(\Lambda(p) - \frac{\alpha p}{n}\sum_{i=0}^{1/p-1} (\Lambda(p)^{1/p})^i) = 0 \\ \iff &\Lambda(p) - \frac{\alpha p}{n}\frac{1-\Lambda(p)}{1-\Lambda(p)^p} = 0 \\ \iff &\Lambda(p)(1-\Lambda(p)^p) = \frac{\alpha p}{n}(1-\Lambda(p)) \end{aligned} \tag{7.7.9}$$

We can derive this expression with respect to p and find:

$$\nabla\Lambda(p)\underbrace{(\overbrace{(1-\Lambda(p)^p)}^{>0} - \Lambda(p)^p\overbrace{\ln(\Lambda(p))}^{<0}p + \frac{\alpha p}{n})}_{>0} = \underbrace{\alpha(1-\Lambda(p))}_{>0} \tag{7.7.10}$$

which yields $\nabla\Lambda(p) > 0$.

Conclusion: The function $\Lambda(p)$ is monotonic which means that the speed of the iEM will be monotonic with the number of individual we pick at each iteration. Thus, faster convergence is attained with the smallest batch size possible, *i.e.,* of size 1.

Incremental SAEM algorithm In the iSAEM only the latent variable whose index has been picked will be simulated. Moreover, it will be simulated by the posterior distribution under the latest model parameter estimate. As a result we have for all $i \in [\![1, n]\!]$:

$$z_i^k \sim p_i(z_i|y_i, \boldsymbol{\theta}^{(\tau_i^k)}) \tag{7.7.11}$$

where $\tau_i^0 = 0$ and for all $k \geq 1$ the index i_k is defined recursively as follows:

$$\tau_i^k = \begin{cases} k-1 & \text{if } i \in I_k \\ \tau_i^{k-1} & \text{otherwise} \end{cases} \tag{7.7.12}$$

In this case the posterior distribution being a Gaussian distribution we can write each latent variable as:

$$z_i^k = \alpha\boldsymbol{\theta}^{(\tau_i^k)} + (1-\alpha)y_j + e^k \ , \tag{7.7.13}$$

where $e^k \sim \mathcal{N}(0, \gamma^2)$. We can now apply our maximization step considering a Monte Carlo batch size M =1 and a mini-batch I_k of size 1:

$$\boldsymbol{\theta}^{(k)} = \hat{\boldsymbol{\theta}}(s^{(k)}) = \frac{1}{n}\sum_{i=1}^{n}(\tilde{S}_i(z_i^k, y_i)|y_i, \boldsymbol{\theta}^{(\tau_i^k)}) = \frac{\alpha}{n}\sum_{i=1}^{n}\boldsymbol{\theta}^{(\tau_i^k)} + (1-\alpha)\bar{y} + \bar{e}^k \ , \tag{7.7.14}$$

where $\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$ and $\bar{e}^k \sim \mathcal{N}(0, \frac{\gamma^2}{n})$. If we define the vector of parameters $\underline{\boldsymbol{\theta}}^k$ as follows:

$$\underline{\boldsymbol{\theta}}^k = \begin{pmatrix} \boldsymbol{\theta}^{(k)} \\ .. \\ \boldsymbol{\theta}^{(k-n+1)} \end{pmatrix} = \rho_1\underline{\boldsymbol{\theta}}^{k-1} + (1-\alpha)\bar{y}\underline{e}_1 + \bar{e}_k\underline{e}_1 \ , \tag{7.7.15}$$

Now if we consider a scheme where not only a single individual is picked at each iteration but rather a mini-batch of size pN (where $p \in [\![0, 1]\!]$) and considering $\delta = \frac{\alpha p}{n}$ and $\bar{y} = 0$

for ease of notations, we can write, in the scalar case:

$$\boldsymbol{\theta}^{(k)} = \delta^{1/p}\boldsymbol{\theta}^{(k-1/p)} + \bar{e}_k \sum_{i=0}^{1/p-1} \delta^i = \delta^{1/p}\boldsymbol{\theta}^{(k-1/p)} + \bar{e}_k \frac{1-\delta^{1/p}}{1-\delta} \tag{7.7.16}$$

We calculate the expectation and the variance of our estimator $\boldsymbol{\theta}^{(k)}$ in the stationary regime:

$$\mathbb{E}[\boldsymbol{\theta}^{(k)}] = \delta^{k/p}\boldsymbol{\theta}^0 \quad \text{and} \quad \text{Var}\,[\boldsymbol{\theta}^{(k)}] = \frac{\gamma^2}{n(1-\delta)^2}\frac{1-\delta^{1/p}}{1+\delta^{1/p}} \tag{7.7.17}$$

Conclusion: With these two expressions, we understand what strategy is best for the choice of the batch size at each iteration. Indeed the bias is small when p is small so one should start with picking one individual first to reduce the bias and the variance is decreasing when p is increasing. Thus, once the bias is reduced one should increase the size of the batch to decrease the variance of the estimator.

Chapter 8

R Tutorial: MLE for Noncontinuous Data Models

Abstract: *This Chapter corresponds to a tutorial on the extension of the* saemix *[Comets et al., 2017] R package. Initially developed to run the SAEM algorithm, introduced in Chapter 6, on continuous data models, we demonstrate in this Chapter, how its extension, useful for estimating population parameters in noncontinuous data models, can be used. The extended package is made available on GitHub:* `https://github.com/belhal/saemix`.

Contents

8.1 Introduction

The R package SAEMIX [Comets et al., 2017] is an implementation in R language [R Development Core Team, 2008] of the Stochastic Approximation Expectation Maximization algorithm developed by Kuhn and Lavielle [2004] and presented in Chapter 6. It is

implemented in the Monolix software available in Matlab and as a standalone software for Windows, MacOS and Linux [Lavielle, 2005].

It performs parameter estimation for nonlinear mixed effects models, goodness of fit plots and model selection (using information criteria such as the AIC or the BIC and testing hypotheses using the Likelihood Ratio Test). The SAEM and its associated package have been used in several application such as agronomy [Makowski and Lavielle, 2006], animal breeding [Jaffrézic et al., 2006] and PK-PD analysis [Bertrand et al., 2009, Lavielle and Mentré, 2007, Samson et al., 2006].

Though, the current version of the SAEMIX R package handles only analytical structural model functions for continuous data models. The present Chapter describes the use of the extended version of the SAEMIX package in R for noncontinuous data models. Two examples (a time-to-event model and a categorical model) are described along with their code implementation.

8.2 Noncontinuous Data Models

As mentioned above, SAEMIX can also be used for noncontinuous data models. Noncontinuous data models include categorical data models [Agresti, 1990, Savic et al., 2011], time-to-event data models [Andersen, 2006, Mbogning et al., 2015], or count data models [Savic et al., 2011].

A categorical outcome y_{ij} takes its value in a set $\{1, \ldots, L\}$ of L categories. Then, the model is defined by the conditional probabilities $(\mathbb{P}(y_{ij} = \ell|\psi_i), 1 \leq \ell \leq L)$, that depend on the vector of individual parameters ψ_i and may be a function of the time t_{ij}.

In a time-to-event data model, the observations are the times at which events occur. An event may be one-off (e.g., death, hardware failure) or repeated (e.g., epileptic seizures, mechanical incidents). To begin with, we consider a model for a one-off event. The survival function $S(t)$ gives the probability that the event happens after time t:

$$S(t) \triangleq \mathbb{P}(T > t) = \exp\left\{-\int_0^t h(u)\mathrm{d}u\right\} , \tag{8.2.1}$$

where h is called the hazard function. In a population approach, we consider a parametric and individual hazard function $h(\cdot, \psi_i)$.

The random variable representing the time-to-event for individual i is typically written T_i and may possibly be right-censored. Then, the observation y_i for individual i is

$$y_i = \begin{cases} T_i & \text{if } T_i \leq \tau_c \\ \text{"}T_i > \tau_c\text{"} & \text{otherwise} , \end{cases} \tag{8.2.2}$$

where τ_c is the censoring time and "$T_i > \tau_c$" is the information that the event occurred after the censoring time.

8.3 A Repeated Time-To-Event Data Model

8.3.1 The model

For repeated event models, times when events occur for individual i are random times $(T_{ij}, 1 \leq j \leq n_i)$ for which conditional survival functions can be defined:

$$\mathbb{P}(T_{ij} > t | T_{i(j-1)} = t_{i(j-1)}) = \exp\left\{ -\int_{t_{i(j-1)}}^{t} h(u, \psi_i) \mathrm{d}u \right\} . \tag{8.3.1}$$

Here, t_{ij} is the observed value of the random time T_{ij}. If the last event is right censored, then the last observation y_{i,n_i} for individual i is the information that the censoring time has been reached "$T_{i,n_i} > \tau_c$". The conditional pdf of $y_i = (y_{ij}, 1 \leq n_i)$ reads (see [Lavielle, 2014] for more details)

$$p_i(y_i|\psi_i) = \exp\left\{ -\int_0^{\tau_c} h(u, \psi_i) \mathrm{d}u \right\} \prod_{j=1}^{n_i - 1} h(t_{ij}, \psi_i) . \tag{8.3.2}$$

8.3.2 Numerical application

In this section, we consider the example developed in Chapter 6. We recall the Weibull model for time-to-event data [Lavielle, 2014, Zhang, 2016]. For individual i, the hazard function of this model is:

$$h(t, \psi_i) = \frac{\beta_i}{\lambda_i} \left(\frac{t}{\lambda_i} \right)^{\beta_i - 1} . \tag{8.3.3}$$

Here, the vector of individual parameters is $\psi_i = (\lambda_i, \beta_i)$. These two parameters are assumed to be independent and lognormally distributed:

$$\log(\lambda_i) \sim \mathcal{N}(\log(\lambda_{\text{pop}}), \omega_\lambda^2) \quad \text{and} \quad \log(\beta_i) \sim \mathcal{N}(\log(\beta_{\text{pop}}), \omega_\beta^2) . \tag{8.3.4}$$

Then, the vector of population parameters is $\theta = (\lambda_{\text{pop}}, \beta_{\text{pop}}, \omega_\lambda, \omega_\beta)$. Repeated events were generated using simulx (mlxR package in R [Lavielle et al., 2019]), for $N = 100$ individuals, using the Weibull model (8.3.3) with $\lambda_{\text{pop}} = 10$, $\omega_\lambda = 0.3$, $\beta_{\text{pop}} = 3$ and $\omega_\beta = 0.3$ and assuming a right censoring time $\tau_c = 20$.

The following code, written in R, is used to run this example. The first step consists in importing the library, the data and initiating the Data Object as follows:

```
library(saemix)
data(tte.saemix)
saemix.data<-saemixData(name.data=tte.saemix,header=TRUE,sep=" ",na=NA, name
    .group=c("id"),name.response=c("y"),name.predictors=c("time","y"), name.X
    =c("time"))
```

We identify, in the object saemix.data, the name of the predictors, the response and the identifier for each individual. Then, the structural model is written in R language and its associated saemix Model Object is created as follows:

```
timetoevent.model<-function(psi,id,xidep) {
  T<-xidep[,1]
  N <- nrow(psi)
  Nj <- length(T)
  censoringtime = 20
  lambda <- psi[id,1]
  beta <- psi[id,2]
  init <- which(T== 0)
  cens <- which(T== censoringtime)
  ind <- setdiff(1:Nj, append(init,cens))
  hazard <- (beta/lambda)*(T/lambda)^(beta-1)
  H <- (T/lambda)^beta
  logpdf <- rep(0,Nj)
  logpdf[cens] <- -H[cens] + H[cens-1]
  logpdf[ind] <- -H[ind] + H[ind-1] + log(hazard[ind])
  return(logpdf) }

saemix.model<-saemixModel(model=timetoevent.model,description="time model",
    type="likelihood", psi0=matrix(c(2,1),ncol=2,byrow=TRUE,dimnames=list(
    NULL, c("lambda","beta"))), transform.par=c(1,1),covariance.model=matrix(
    c(1,0,0,1),ncol=2, byrow=TRUE))
```

We note in this code snippet that the model function, called timetoevent.model, defines the log pdf, as written in (8.3.2), of the time-to-event model. The additional argument type ="likelihood", allows us to run the SAEM on such noncontinuous model (type="structural" is the value of that argument for continuous data models. In that case, the model function returns the structural model f, as defined in Chapter 6). Finally, we define the list of the SAEM hyperparameters, such as the number of iterations or MCMC chains, and run the algorithm as follows:

```
saemix.options<-list(map=F,fim=F,ll.is=F, nb.chains = 1, nbiter.saemix = c
    (200,100),displayProgress=TRUE,save.graphs=FALSE)
saemix.fit<-saemix(model,saemix.data,saemix.options)
```

Figure 8.1 shows the convergence of the population parameters for this example. The results are summed up in the following table:

```
-----------------------------------------------------------
```

```
-----                    Results                    -----

-------------------------------------------------------

--------------------  Fixed effects  --------------------

-------------------------------------------------------

        Parameter Estimate
[1,] lambda      5.0
[2,] beta        2.8

-------------------------------------------------------

-------------  Variance of random effects  -------------

-------------------------------------------------------

        Parameter       Estimate
lambda omega2.lambda 0.039
beta    omega2.beta    0.921

-------------------------------------------------------

-------  Correlation matrix of random effects  -------

-------------------------------------------------------

                 omega2.lambda omega2.beta
omega2.lambda 1                0
omega2.beta    0                1
```

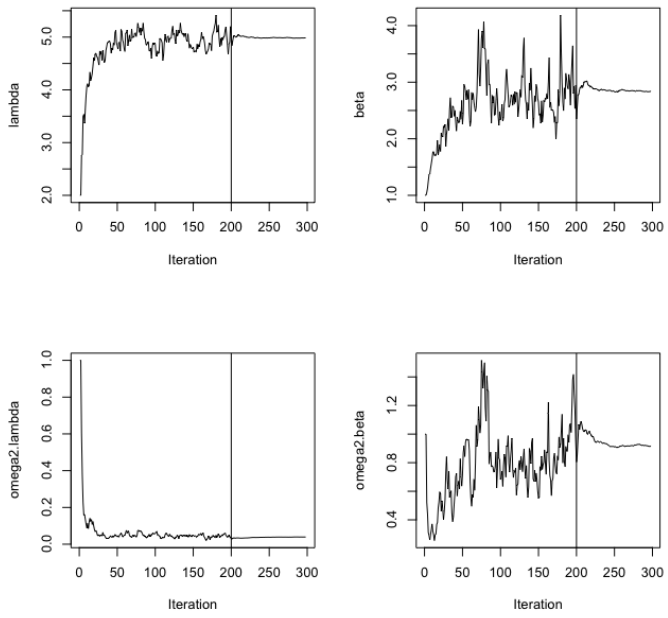

Figure 8.1 – Time-to-event data modelling: convergence of the population parameters (λ, β) and the random effects $(\omega^2_\lambda, \omega^2_\beta)$.

8.4 A Categorical Data Model with Regression Variables

8.4.1 The model

Assume now that the observed data takes its values in a fixed and finite set of nominal categories $\{c_1, c_2, \ldots, c_K\}$. Considering the observations $(y_{ij},\, 1 \leq j \leq n_i)$ for any individual i as a sequence of conditionally independent random variables, the model is completely defined by the probability mass functions $\mathbb{P}(y_{ij} = c_k|\psi_i)$ for $k = 1, \ldots, K$ and $1 \leq j \leq n_i$. For a given (i, j), the sum of the K probabilities is 1, so in fact only $K - 1$ of them need to be defined. In the most general way possible, any model can be considered so long as it defines a probability distribution, i.e., for each k, $\mathbb{P}(y_{ij} = c_k|\psi_i) \in [0, 1]$, and $\sum_{k=1}^{K} \mathbb{P}(y_{ij} = c_k|\psi_i) = 1$. Ordinal data further assume that the categories are ordered, i.e., there exists an order $\prec$ such that

$$c_1 \prec c_2, \prec \ldots \prec c_K$$

.

We can think, for instance, of levels of pain ($low \prec moderate \prec severe$) or scores on a discrete scale, e.g., from 1 to 10. Instead of defining the probabilities of each category, it may be convenient to define the cumulative probabilities $\mathbb{P}(y_{ij} \preceq c_k|\psi_i)$ for $k = 1, \ldots, K-1$, or in the other direction: $\mathbb{P}(y_{ij} \succeq c_k|\psi_i)$ for $k = 2, \ldots, K$. Any model is possible as long as it defines a probability distribution, i.e., it satisfies

$$0 \leq \mathbb{P}(y_{ij} \prec c_1|\psi_i) \leq \mathbb{P}(y_{ij} \prec c_2|\psi_i) \leq \cdots \leq \mathbb{P}(y_{ij} \prec c_K|\psi_i) = 1$$

It is possible to introduce dependence between observations from the same individual by assuming that $(y_{ij},\, j = 1, 2, \ldots, n_i)$ forms a Markov chain. For instance, a Markov chain with memory 1 assumes that all that is required from the past to determine the distribution of y_{ij} is the value of the previous observation $y_{i,j-1}$., i.e., for all $k = 1, 2, \ldots, K$,

$$\mathbb{P}(y_{ij} = c_k|y_{i,j-1}, y_{i,j-2}, y_{i,j-3}, \psi_i) = \mathbb{P}(y_{ij} = c_k|y_{i,j-1}, \psi_i)$$

8.4.2 Numerical application

In this example, observations are ordinal data that take their values in $\{0, 1\}$ Odds ratio are used in this example to define the model

$$\text{logit}(\mathbb{P}(y_{ij} = k)) = \log \frac{\mathbb{P}(y_{ij} = k)}{1 - \mathbb{P}(y_{ij} = k)}$$

where y_{ij} denotes the j-th observation for the i-th individual and:

$$\text{logit}(\mathbb{P}(y_{ij}=0)) = \theta_{i,1} + \theta_{i,2}\,\text{Time}_{\text{ij}} + \theta_{i,3}\,\text{Dose}_{\text{i}} \tag{8.4.1}$$

where Dose and Time are the two regression variables.

Here, the vector of individual parameters is $\psi_i = (\theta_{i,1}, \theta_{i,2}, \theta_{i,3})$. These three parameters are assumed to be independent and normally distributed:

$$\theta_{i,1} \sim \mathcal{N}(\theta_{\text{pop},1}, \omega_1^2)\ , \theta_{i,1} \sim \mathcal{N}(\theta_{\text{pop},2}, \omega_2^2)\ , \theta_{i,1} \sim \mathcal{N}(\theta_{\text{pop},3}, \omega_3^2) \tag{8.4.2}$$

Then, the vector of population parameters is $\theta = (\theta_{\text{pop},1}, \theta_{\text{pop},2}, \theta_{\text{pop},1}, \omega_1, \omega_2, \omega_3)$.

Data simulation: Data is generated using $N = 300$ and for all $i \in [\![1, n]\!]$, $n_i = 15$. For all $i \in [\![1, n]\!]$ and $j \in [\![n_i]\!]$, we take $d_{ij,1} = 1$, $d_{ij,2} = -20 + (j-1) * 5$ and for $i \in [\![1, n]\!]$ $d_{ij,3} = 10\lceil 3i/N \rceil$. The data is generated using the following values for the fixed and random effects ($\theta_{\text{pop},1} = -4, \theta_{\text{pop},2} = -0.5, \theta_{\text{pop},} = 1, \omega_1 = 0.3, \omega_2 = 0.2, \omega_3 = 0.2$). Here is a sample code on how to generate such data using the mlxR [Lavielle et al., 2019] package:

```
library("mlxR")
catModel <- inlineModel(
"[LONGITUDINAL]
input =  {beta0,gamma0,delta0, dose}
dose = {use=regressor}
EQUATION:
lm0 = beta0+gamma0*t + delta0*dose
D = exp(lm0)+1
p0 = exp(lm0)/D
p1 = 1/D

DEFINITION:
y = {type=categorical, categories={0, 1},
     P(y=0)=p0,
     P(y=1)=p1}
[INDIVIDUAL]
input={beta0_pop, o_beta0,
       gamma0_pop, o_gamma0,
       delta0_pop, o_delta0}
DEFINITION:
beta0  ={distribution=normal, prediction=beta0_pop,  sd=o_beta0}
gamma0  ={distribution=normal, prediction=gamma0_pop,  sd=o_gamma0}
delta0  ={distribution=normal, prediction=delta0_pop,  sd=o_delta0} ")

nobs = 15
tobs<- seq(-20, 50, by=nobs)
reg1 <- list(name='dose',
             time=tobs,
             value=10*(tobs>0))
```

```
30
31 reg2 <- list(name='dose',
32               time=tobs,
33               value=20*(tobs>0))
34
35 reg3 <- list(name='dose',
36               time=tobs,
37               value=30*(tobs>0))
38
39 out  <- list(name='y', time=tobs)
40 N  <- 100
41 p <- c(beta0_pop=-4, o_beta0=0.3,
42         gamma0_pop= -0.5, o_gamma0=0.2,
43         delta0_pop=1, o_delta0=0.2)
44
45 g1 <- list(size=N,regressor = reg1)
46 g2 <- list(size=N,regressor = reg2)
47 g3 <- list(size=N,regressor = reg3)
48 g <- list(g1,g2,g3)
49 res <- simulx(model=catModel,output=out, group=g,parameter=p)
50 plot1 <- catplotmlx(res$y)
```

Figure 8.3 shows the probability for each of the three subgroups:

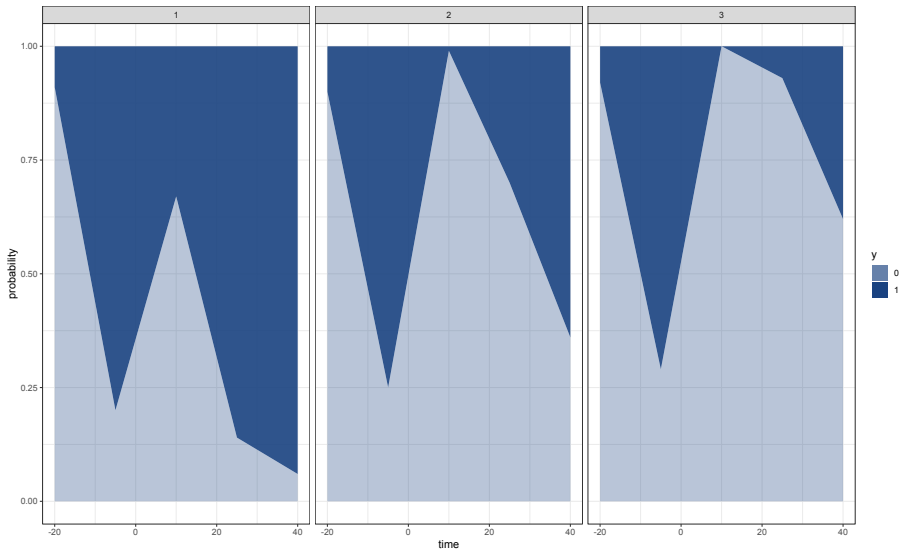

Figure 8.2 – Categorical data modelling: Probabilities $\mathbb{P}(y = 0)$ and $\mathbb{P}(y = 1)$ for each of the three subgroups.

The model is implemented in R as follows:

```
1 cat.model<-function(psi,id,xidep) {
2 level<-xidep[,1]
3 dose<-xidep[,2]
4 time<-xidep[,3]
```

```
th1 <- psi[id,1]
th2 <- psi[id,2]
delta0 <- psi[id,3]
lm0 <- th1+th2*time + delta0*dose
D <- exp(lm0)+1
P0 <- exp(lm0)/D
P1 <- 1/D

P.obs = (level==0)*P0+(level==1)*P1
return(P.obs) }
```

Then, the following code was used in R to run the SAEM algorithm on this example:

```
saemix.model<-saemixModel(model=cat.model,description="cat model",type="
    likelihood", psi0=matrix(c(2,1,2),ncol=3,byrow=TRUE,dimnames=list(NULL,c(
    "th1","th2","th3"))), transform.par=c(0,1,1),covariance.model=matrix(c
    (1,0,0,0,1,0,0,0,1),ncol=3,byrow=TRUE),omega.init=matrix(c
    (2,0,0,0,1,0,0,0,1),ncol=3,byrow=TRUE),error.model="constant")

K1 = 500
K2 = 100
#Saemix Run
options<-list(seed=39546,map=F,fim=F,ll.is=F,
  nbiter.mcmc = c(2,2,2), nbiter.saemix = c(K1,K2),nbiter.sa=0,
  displayProgress=TRUE,save.graphs=FALSE,nbiter.burn =0)
saemix.fit<-saemix(saemix.model,saemix.data,options)
```

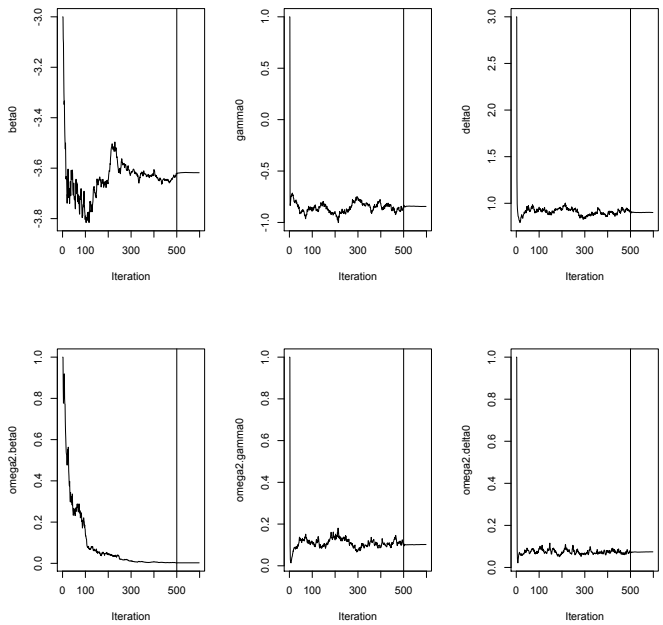

Figure 8.3 – Categorical data modelling: convergence of the population parameters $(\beta_0, \gamma_0, \delta_0)$ and the random effects $(\omega^2_{\beta_0}, \omega^2_{\gamma_0}, \omega^2_{\delta_0})$.

Chapter 9

Conclusion

9.1 Summary of the Thesis

In thesis, we focused on optimization for non-convex objective functions for a particular class of models: latent data models. For either composite function, expected risk or incomplete likelihood functions, we address the data fitting problem with novel algorithms and their corresponding finite-time and asymptotic analyses.

In our first main contribution, we provided MISSO, a general incremental optimization framework for composite objective functions. The scheme is motivated by the use of flexible surrogates illustrated through examples such as variational inference and Monte Carlo EM. We derived non-asymptotic and asymptotic convergence of the iterates and utilized our method to train a logistic regression on the TraumaBase dataset and a Bayesian neural network on the MNIST. This unifying framework is shown to achieve an ϵ-stationary point in $\mathcal{O}(n/\epsilon)$ iterations, yielding at the same time the first non-asymptotic rate for MISO (see [Mairal, 2015a]) to optimize non-convex objective functions.

In our second main contribution, a simple Stochastic Approximation scheme is analyzed under mild assumptions. The main result we derived shows that the SA scheme finds an $\mathcal{O}(c_0 + \log n/\sqrt{n})$ quasi-stationary point within n iterations where the drift term of the algorithm is non necessarily a gradient and can be a biased estimator, of bias c_0, of the mean-field. We applied our results on the online EM algorithm and the Policy gradient algorithm for average reward over infinite horizon with rigorous verification of the assumptions.

In our third main contribution, we studied several incremental variants of the EM algorithm. We focused on the inference of latent variable models with exponential family distribution and analyze the convergence of several stochastic EM methods. We established these analyses based on two complimentary views, one that interprets incremental

EM method as an incremental Majorization-Minimization method, and one that interprets variance reduced and fast variants as scaled gradient,*i.e.,* non-gradients, methods. Numerical applications illustrate the advantages of those latter methods.

In our fourth main contribution, we considered the SAEM algorithm to train mixed effects models. The specificity of the model implies simulating individual parameters at each iteration from their intractable posterior distributions. When this sampling step is performed using an MCMC procedure, we provided in that contribution, an efficient independent Metropolis Hastings proposal to sample from this target. Based on a first-order Taylor expansion and the Laplace approximation of the incomplete log-likelihood, our proposal is a simple Gaussian distribution centered around the mode of the target. A through empirical study, presented in this contribution, highlights the advantages of the proposed sampler as well as its virtue when embedded in a maximum likelihood estimation algorithm such as the SAEM.

In our fifth main contribution, we derived an incremental variant of the SAEM algorithm. Asymptotic convergence result was established and observed through numerical applications. A pharmacokinetics model was trained using this incremental algorithm and exhibited faster convergence of the parameters with a specific mini-batch size and sampling strategy.

In our sixth main contribution, we developed an extension of an R package, called SAEMIX, to perform maximum likelihood estimation on noncontinuous data models. We provided two examples, Categorical data and Time-to-event data models, along with their implementation codes.

9.2 Perspectives

This thesis aims at participating to the general efforts towards understanding how complex models are trained on large datasets. Yet, much more challenges are left out and are worth studying in the future. We give some of our thoughts in the following non exhaustive list:

- A question that arises while using incremental algorithms is the choice of the indices at each iteration. We considered examples with uniform sampling strategy (see Chapter 3-5-7), yet optimal ones are worth deriving. Efforts in that direction include [Roux et al., 2012] where the index of the gradient computed at a given iteration corresponds to the biggest, in norm, gradient term. Of course this strategy is computationally involved but is intuitively optimal for gradient algorithms. Another recent work, [Horváth and Richtárik, 2018], provides optimal rates for SVRG and SAGA under arbitrary, *i.e.,* non necessarily uniform, sampling strategies.
- Another straightforward question regards the optimal mini-batch size of stochastic

and incremental algorithms. Indeed, mini-batch of size 1 are not always optimal as developed in [Gower et al., 2019] where a *variance-cost* trade off has been highlighted along with some results on the choice of the mini-batch size for SGD.

- Besides a possible *variance-cost* trade off, we observed Chapter 3 and Chapter 7, a *bias-variance* trade off due to the Monte Carlo integration of the quantities of interest. A thorough study of the effect of this approximation on the variance of the estimator in the incremental setting would be interesting. This study would intuitively consider the interplay between the Monte Carlo batch and the mini-batch of indices drawn at each iteration. We remind that these types of algorithms involve two levels of stochasticity with two batch sizes as hyper parameters.

- We recall that a complexity of $\mathcal{O}(n/\epsilon)$ was found for the MISO (and MISSO) method. Yet an interesting research direction consists in finding tight upper bounds of this incremental scheme for specific class of surrogate functions in the non-convex setting. For instance Qian et al. [2019] provide an iteration complexity of $\mathcal{O}(n^{2/3}/\epsilon)$ for the MISO method using quadratic surrogates.

- Finally, all those incremental methods obviously trigger the question of storage and computation. While methods like SVRG require less storage than incremental methods such as SAG or SAGA, there must be a particular focus on how to deal with storing values when the data size is big. Regarding computation resources, many works focus on distributed first-order optimization procedures, *i.e.,* dispatching the computation to a cluster of machines, instead of just one. Many challenges can be addressed such as whether or not using a parallel or asynchronous method, or even using a centralized or decentralized architecture.

Scientific Production

Proceedings of international peer-reviewed conferences

On the Global Convergence of (Fast) Incremental Expectation Maximization Methods, Belhal Karimi, Marc Lavielle, Eric Moulines, Hoi-To Wai, Advances in Neural Information Processing Systems, NeurIPS 2019.

Non-asymptotic Analysis of Biased Stochastic Approximation Scheme, Belhal Karimi, Blazej Miasojedow, Eric Moulines and Hoi-To Wai, Proceedings of Conference on Learning Theory, COLT 2019.

Efficient Metropolis-Hastings sampling for nonlinear mixed effects models, Belhal Karimi and Marc Lavielle, Proceedings of BAYSM 2018.

Articles in peer-reviewed journals

f-SAEM: A fast Stochastic Approximation of the EM algorithm, Belhal Karimi, Marc Lavielle and Eric Moulines, Computational Statistics and Data Analysis (CSDA), 2019.

Preprints

A Doubly Stochastic Surrogate Optimization Scheme for Non-convex Finite-sum Problems, Belhal Karimi, Eric Moulines and Hoi-To Wai.

Software

R package, extension of *saemix* (2019).

Awards

Visiting Student Researcher Grant from the Jacques Hadamard Foundation, HSE-Samsung AI Lab in Moscow (RUSSIA) with Dr. Dmitry Vetrov (2 months).

Student award, 32nd Conference on Learning Theory (2019).

Bibliography

Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. (2015). TensorFlow: Large-scale machine learning on heterogeneous systems. Software available from tensorflow.org.

Abbasi-Yadkori, Y., Lazic, N., and Szepesvari, C. (2018). Regret bounds for model-free linear quadratic control. *arXiv preprint arXiv:1804.06021*.

Ablin, P., Gramfort, A., Cardoso, J.-F., and Bach, F. (2018). EM algorithms for ICA. *arXiv preprint arXiv:1805.10054*.

Agarwal, A. and Bottou, L. (2014). A lower bound for the optimization of finite sums. *arXiv preprint arXiv:1410.0723*.

Agarwal, A. and Duchi, J. C. (2013). The generalization ability of online algorithms for dependent data. *IEEE Transactions on Information Theory*, 59(1):573–587.

Agarwal, N., Allen-Zhu, Z., Bullins, B., Hazan, E., and Ma, T. (2017). Finding approximate local minima faster than gradient descent. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing*, pages 1195–1199. ACM.

Agresti, A. (1990). *Categorical data analysis*. A Wiley-Interscience publication. Wiley, New York.

Allassonnière, S. and Chevallier, J. (2019). A New Class of EM Algorithms. Escaping Local Minima and Handling Intractable Sampling. working paper or preprint.

Allassonniere, S. and Kuhn, E. (2013). Convergent Stochastic Expectation Maximization algorithm with efficient sampling in high dimension. Application to deformable template model estimation. *arXiv preprint arXiv:1207.5938*.

Allen-Zhu, Z. and Hazan, E. (2016). Variance reduction for faster non-convex optimization. In *International Conference on Machine Learning*, pages 699–707.

Andersen, P. K. (2006). Survival Analysis. *Wiley Reference Series in Biostatistics*.

Andrieu, C., Doucet, A., and Holenstein, R. (2010). Particle markov chain monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 72(3):269–342.

Andrieu, C., Moulines, É., et al. (2006). On the ergodicity properties of some adaptive mcmc algorithms. *The Annals of Applied Probability*, 16(3):1462–1505.

Andrieu, C., Roberts, G. O., et al. (2009). The pseudo-marginal approach for efficient monte carlo computations. *The Annals of Statistics*, 37(2):697–725.

Andrieu, C. and Thoms, J. (2008). A tutorial on adaptive mcmc. *Statistics and computing*, 18(4):343–373.

Atchadé, Y. F., Rosenthal, J. S., et al. (2005). On adaptive markov chain monte carlo algorithms. *Bernoulli*, 11(5):815–828.

Balakrishnan, S., Wainwright, M. J., and Yu, B. (2017). Statistical guarantees for the EM algorithm: From population to sample-based analysis. *Ann. Statist.*, 45(1):77–120.

Baxter, J. and Bartlett, P. L. (2001). Infinite-horizon policy-gradient estimation. *Journal of Artificial Intelligence Research*, 15:319–350.

Beal, S. and Sheiner, L. (1980). The NONMEM system. *The American Statistician*, 34(2):118–119.

Benveniste, A., Priouret, P., and Métivier, M. (1990). *Adaptive Algorithms and Stochastic Approximations*. Springer-Verlag, Berlin, Heidelberg.

Bertrand, J., Comets, E., Laffont, C. M., Chenel, M., and Mentré, F. (2009). Pharmacogenetics and population pharmacokinetics: impact of the design on three tests using the saem algorithm. *Journal of pharmacokinetics and pharmacodynamics*, 36(4):317–339.

Bertsekas, D. P. (1999). *Nonlinear programming*. Athena scientific Belmont.

Bertsekas, D. P. (2011). Incremental gradient, subgradient, and proximal methods for convex optimization: A survey. *Optimization for Machine Learning*, 2010(1-38):3.

Betancourt, M. (2017). A conceptual introduction to Hamiltonian Monte Carlo. *arXiv preprint arXiv:1701.02434*.

Bhandari, J., Russo, D., and Singal, R. (2018). A finite time analysis of temporal difference learning with linear function approximation. In *Conference On Learning Theory*, pages 1691–1692.

Bishop, C. M. (2006). *Pattern recognition and machine learning*. springer.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017a). Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877.

Blei, D. M., Kucukelbir, A., and McAuliffe, J. D. (2017b). Variational Inference: A Review for Statisticians. *Journal of the American statistical Association*, 112(518):859–877.

Blundell, C., Cornebise, J., Kavukcuoglu, K., and Wierstra, D. (2015). Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622.

Borkar, V. S. (1997). Stochastic approximation with two time scales. *Systems & Control Letters*, 29(5):291–294.

Borkar, V. S. (2009). *Stochastic approximation: a dynamical systems viewpoint*, volume 48. Springer.

Bottou, L. (1991). Stochastic gradient learning in neural networks. *Proceedings of Neuro-Nımes*, 91(8):12.

Bottou, L. (1998). Online learning and stochastic approximations. *On-line learning in neural networks*, 17(9):142.

Bottou, L. and Bousquet, O. (2008). The tradeoffs of large scale learning. In Platt, J. C., Koller, D., Singer, Y., and Roweis, S. T., editors, *Advances in Neural Information Processing Systems 20*, pages 161–168. Curran Associates, Inc.

Bottou, L., Curtis, F. E., and Nocedal, J. (2018). Optimization methods for large-scale machine learning. *SIAM Review*, 60(2):223–311.

Bottou, L. and Le Cun, Y. (2005). On-line learning for very large data sets. *Applied stochastic models in business and industry*, 21(2):137–151.

Boucheron, S., Lugosi, G., and Massart, P. (2013). *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press.

Boyd, S. and Vandenberghe, L. (2004). *Convex optimization*. Cambridge university press.

Brocker, T., Lander, L., et al. (1975). *Differentiable germs and catastrophes*, volume 17. Cambridge University Press.

Brooks, S., Gelman, A., Jones, G., and Meng, X.-L. (2011). *Handbook of markov chain monte carlo*. CRC press.

Brosse, N., Durmus, A., Moulines, É., and Sabanis, S. (2017). The tamed unadjusted langevin algorithm. *arXiv preprint arXiv:1710.05559*.

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

Carmon, Y., Duchi, J. C., Hinder, O., and Sidford, A. (2017). Convex until proven guilty: Dimension-free acceleration of gradient descent on non-convex functions. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 654–663. JMLR. org.

Carpenter, B., Gelman, A., Hoffman, M., Lee, D., Ben, G., Betancourt, M., Brubaker, M., Guo, J., Li, P., and Riddell, A. (2017). Stan: A probabilistic programming language. *Journal of Statistical Software*, 76(1).

Chan, P. L. S., Jacqmin, P., Lavielle, M., McFadyen, L., and Weatherley, B. (2011). The use of the SAEM algorithm in MONOLIX software for estimation of population pharmacokinetic-pharmacodynamic-viral dynamics parameters of maraviroc in asymptomatic HIV subjects. *Journal of Pharmacokinetics and Pharmacodynamics*, 38(1):41–61.

Chen, J., Zhu, J., Teh, Y. W., and Zhang, T. (2018). Stochastic Expectation Maximization with variance reduction. In *Advances in Neural Information Processing Systems*, pages 7978–7988.

Comets, E., Lavenu, A., and Lavielle, M. (2017). Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software*, 80(3):1–42.

Conn, A. R., Gould, N., Sartenaer, A., and Toint, P. L. (1993). Global convergence of a class of trust region algorithms for optimization using inexact projections on convex constraints. *SIAM Journal on Optimization*, 3(1):164–221.

Csiszár, I. and Tusnády, G. (1984). Information geometry and alternating minimization procedures. *Statist. Decisions*, suppl. 1:205–237. Recent results in estimation theory and related topics.

Dalal, G., Szörényi, B., Thoppe, G., and Mannor, S. (2018a). Finite sample analyses for td (0) with function approximation. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Dalal, G., Szorenyi, B., Thoppe, G., and Mannor, S. (2018b). Finite sample analysis of two-timescale stochastic approximation with applications to reinforcement learning. In *Conference On Learning Theory*.

Davidian, M. (2017). *Nonlinear models for repeated measurement data*. Routledge.

de Freitas, N., Højen-Sørensen, P., Jordan, M. I., and Russell, S. (2001). Variational mcmc. *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, pages 120–127.

Defazio, A., Bach, F., and Lacoste-Julien, S. (2014). Saga: A fast incremental gradient method with support for non-strongly convex composite objectives. In *Advances in neural information processing systems*, pages 1646–1654.

Degris, T., White, M., and Sutton, R. S. (2012). Off-policy actor-critic. *arXiv preprint arXiv:1205.4839*.

Delyon, B., Lavielle, M., and Moulines, E. (1999a). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128.

Delyon, B., Lavielle, M., and Moulines, E. (1999b). Convergence of a stochastic approximation version of the EM algorithm. *Ann. Statist.*, 27(1):94–128.

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 (methodological)*, pages 1–38.

Dillon, J. V., Langmore, I., Tran, D., Brevdo, E., Vasudevan, S., Moore, D., Patton, B., Alemi, A., Hoffman, M. D., and Saurous, R. A. (2017). Tensorflow distributions. *CoRR*, abs/1711.10604.

Donnet, S. and Samson, A. (2013). Using pmcmc in EM algorithm for stochastic mixed models: theoretical and practical issues. *Journal de la Société Française de Statistique*, 155(1):49–72.

Douc, R., Moulines, E., and Stoffer, D. (2014). *Nonlinear Time Series: Theory, Methods and Applications with R examples*. Chapman and Hall/CRC.

Doucet, A., Godsill, S., and Andrieu, C. (2000). On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing*, 10(3):197–208.

Doukhan, P., Massart, P., and Rio, E. (1995). Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 393–427.

Duchi, J. C., Agarwal, A., Johansson, M., and Jordan, M. I. (2012). Ergodic mirror descent. *SIAM Journal on Optimization*, 22(4):1549–1578.

Fang, C., Li, C. J., Lin, Z., and Zhang, T. (2018). Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. In *Advances in Neural Information Processing Systems*, pages 687–697.

Fazel, M., Ge, R., Kakade, S., and Mesbahi, M. (2018). Global convergence of policy gradient methods for the linear quadratic regulator. In *International Conference on Machine Learning*, pages 1466–1475.

Fisher, R. A. (1925). Theory of statistical estimation. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 22, pages 700–725. Cambridge University Press.

Fort, G., Moulines, E., and Priouret, P. (2011). Convergence of adaptive and interacting Markov chain monte carlo algorithms. *The Annals of Statistics*, 39(6):3262–3289.

Fraley, C. and Raftery, A. E. (2007). Bayesian regularization for normal mixture estimation and model-based clustering. *Journal of classification*, 24(2):155–181.

Ghadimi, S. and Lan, G. (2013). Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368.

Ghadimi, S., Lan, G., and Zhang, H. (2016). Mini-batch stochastic approximation methods for nonconvex stochastic composite optimization. *Mathematical Programming*, 155(1-2):267–305.

Ghahramani, Z. (2015). Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459. On Probabilistic models.

Gower, R. M., Loizou, N., Qian, X., Sailanbayev, A., Shulgin, E., and Richtárik, P. (2019). SGD: general analysis and improved rates. *CoRR*, abs/1901.09401.

Griewank, A. and Walther, A. (2008). *Evaluating derivatives: principles and techniques of algorithmic differentiation*, volume 105. Siam.

Gunawardana, A. and Byrne, W. (2005). Convergence theorems for generalized alternating minimization procedures. *Journal of Machine Learning Research*, 6:2049–2073.

Haario, H., Saksman, E., Tamminen, J., et al. (2001). An adaptive metropolis algorithm. *Bernoulli*, 7(2):223–242.

Hall, P. and Heyde, C. C. (2014). *Martingale limit theory and its application*. Academic press.

Hinton, G. E., Osindero, S., and Teh, Y.-W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, 18(7):1527–1554.

Hoffman, M. D. and Gelman, A. (2014). The No-U-turn sampler: adaptively setting path lengths in Hamiltonian Monte Carlo. *Journal of Machine Learning Research*, 15(1):1593–1623.

Hofmann, T. (1999). Probabilistic latent semantic indexing. In *Proceedings of the 22Nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '99, pages 50–57, New York, NY, USA. ACM.

Horváth, S. and Richtárik, P. (2018). Nonconvex variance reduced optimization with arbitrary sampling. *arXiv preprint arXiv:1809.04146*.

J. Reddi, S., Sra, S., Poczos, B., and Smola, A. J. (2016). Proximal stochastic methods for nonsmooth nonconvex finite-sum optimization. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 1145–1153. Curran Associates, Inc.

Jaakkola, T., Jordan, M. I., and Singh, S. P. (1994). Convergence of stochastic iterative dynamic programming algorithms. In *Advances in Neural Information Processing Systems*, pages 703–710.

Jaffrézic, F., Meza, C., Lavielle, M., and Foulley, J.-L. (2006). Genetic analysis of growth curves using the saem algorithm. *Genetics Selection Evolution*, 38(6):583.

Jiang, W., Josse, J., and Lavielle, M. (2018). Logistic regression with missing covariates–parameter estimation, model selection and prediction. arXiv.

Johnson, R. and Zhang, T. (2013). Accelerating stochastic gradient descent using predictive variance reduction. In *Advances in neural information processing systems*, pages 315–323.

Jordan, M. I., Ghahramani, Z., Jaakkola, T. S., and Saul, L. K. (1999). An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233.

Kalman, R. E. and Bertram, J. E. (1960). Control system analysis and design via the "second method" of lyapunov: I—continuous-time systems. *Journal of Basic Engineering*, 82(2):371–393.

Karimi, B. and Lavielle, M. (2018). Efficient Metropolis-Hastings sampling for nonlinear mixed effects models. *Proceedings of BAYSM 2018*.

Karimi, B., Lavielle, M., and Moulines, E. (2020). f-saem: A fast stochastic approximation of the em algorithm for nonlinear mixed effects models. *Computational Statistics & Data Analysis*, 141:123–138.

Karimi, B., Miasojedow, B., Moulines, E., and Wai, H.-T. (2019a). Non-asymptotic analysis of biased stochastic approximation scheme. In Beygelzimer, A. and Hsu, D., editors, *Proceedings of the Thirty-Second Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 1944–1974, Phoenix, USA. PMLR.

Karimi, B., Wai, H.-T., and Moulines, E. (2019b). A doubly stochastic surrogate optimization scheme for non-convex finite-sum problems. *Submitted paper*.

Karimi, B., Wai, H.-T., Moulines, E., and Lavielle, M. (2019c). On the global convergence of (fast) incremental expectation maximization methods. In *Advances in Neural Information Processing Systems*.

Kingma, D. P. and Ba, J. (2015). Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kingma, D. P. and Welling, M. (2014). Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*.

Konda, V. R. and Tsitsiklis, J. N. (2003). On actor-critic algorithms. *SIAM journal on Control and Optimization*, 42(4):1143–1166.

Kucukelbir, A., Ranganath, R., Gelman, A., and Blei, D. (2015). Automatic variational inference in stan. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 568–576. Curran Associates, Inc.

Kuhn, E. and Lavielle, M. (2004). Coupling a stochastic approximation version of EM with an MCMC procedure. *ESAIM: Probability and Statistics*, 8:115–131.

Kushner, H. and Yin, G. G. (2003). *Stochastic approximation and recursive algorithms and applications*, volume 35. Springer Science & Business Media.

Kushner, H. J. and Clark, D. S. (2012). *Stochastic approximation methods for constrained and unconstrained systems*, volume 26. Springer Science & Business Media.

Lakshminarayanan, C. and Szepesvari, C. (2018). Linear stochastic approximation: How far does constant step-size and iterate averaging go? In *International Conference on Artificial Intelligence and Statistics*, pages 1347–1355.

Lange, K. (2016). *MM Optimization Algorithms*. SIAM-Society for Industrial and Applied Mathematics, USA.

Lavielle, M. (1995). A stochastic algorithm for parametric and non-parametric estimation in the case of incomplete data. *Signal Processing*, 42(1):3–17.

Lavielle, M. (2005). Monolix (modèles non linéaires à effets mixtes). *MONOLIX group, Orsay, France*.

Lavielle, M. (2014). *Mixed effects models for the population approach: models, tasks, methods and tools*. CRC press.

Lavielle, M., Ilinca, E., and Kuate, R. (2019). *mlxR: Simulation of Longitudinal Data*. R package version 4.0.0.

Lavielle, M. and Mentré, F. (2007). Estimation of population pharmacokinetic parameters of saquinavir in hiv patients with the monolix software. *Journal of pharmacokinetics and pharmacodynamics*, 34(2):229–249.

Lavielle, M. and Ribba, B. (2016). Enhanced method for diagnosing pharmacometric models: random sampling from conditional distributions. *Pharmaceutical research*, 33(12):2979–2988.

LeCun, Y. (1998). The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*.

LeCun, Y., Bottou, L., Bengio, Y., Haffner, P., et al. (1998). Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324.

Li, Y. and Gal, Y. (2017). Dropout inference in bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2052–2061. JMLR. org.

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.

Mairal, J. (2015a). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855.

Mairal, J. (2015b). Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM J. Optim.*, 25(2):829–855.

Makowski, D. and Lavielle, M. (2006). Using saem to estimate parameters of models of response to applied fertilizer. *Journal of Agricultural, Biological, and Environmental Statistics*, 11(1):45–60.

Mbogning, C., Bleakley, K., and Lavielle, M. (2015). Joint modeling of longitudinal and repeated time-to-event data using nonlinear mixed-effects models and the SAEM algorithm. *Journal of Statistical Computation and Simulation*, 85(8):1512–1528.

McLachlan, G. and Krishnan, T. (2007). *The EM algorithm and extensions*, volume 382. John Wiley & Sons.

McLachlan, G. J. and Krishnan, T. (2008). *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition.

Medelyan, O. (2009). *Human-competitive automatic topic indexing*. PhD thesis, The University of Waikato.

Mengersen, K. L. and Tweedie, R. L. (1996). Rates of convergence of the hastings and metropolis algorithms. *Ann. Statist.*, 24(1):101–121.

Metropolis, N., Rosenbluth, A. W., Rosenbluth, M. N., Teller, A. H., and Teller, E. (1953). Equation of state calculations by fast computing machines. *The Journal of Chemical Physics*, 21(6):1087–1092.

Meyn, S. P. and Tweedie, R. L. (2012). *Markov chains and stochastic stability*. Springer Science & Business Media.

Migon, H., Gamerman, D., and Louzada, F. (2014). *Statistical Inference: An Integrated Approach, Second Edition*. Chapman & Hall/CRC Texts in Statistical Science. Taylor & Francis.

Moulines, E. and Bach, F. R. (2011). Non-asymptotic analysis of stochastic approximation

algorithms for machine learning. In *Advances in Neural Information Processing Systems*, pages 451–459.

Neal, R. M. (2012). *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media.

Neal, R. M. et al. (2011). Mcmc using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, 2(11):2.

Neal, R. M. and Hinton, G. E. (1998). A view of the EM algorithm that justifies incremental, sparse, and other variants. In *Learning in graphical models*, pages 355–368. Springer.

Nemirovsky A.S., A. S. and IUdin, D. B. (1983). *Problem complexity and method efficiency in optimization /*. Wiley„ Chichester ;. "A Wiley-Interscience publication.".

Nesterov, Y. (2004). *Introductory Lectures on Convex Optimization: A Basic Course*. Springer Publishing Company, Incorporated, 1 edition.

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

O'Reilly, R. A. and Aggeler, P. M. (1968). Studies on coumarin anticoagulant drugs initiation of warfarin therapy without a loading dose. *Circulation*, 38(1):169–177.

Paisley, J., Blei, D., and Jordan, M. (2012). Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress.

Papini, M., Binaghi, D., Canonaco, G., Pirotta, M., and Restelli, M. (2018). Stochastic variance-reduced policy gradient. In Dy, J. and Krause, A., editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 4026–4035, Stockholmsmassan, Stockholm Sweden. PMLR.

Pav, S. E. (2016). *Madness: a package for Multivariate Automatic Differentiation*. R package version 0.2.6.

Peters, J. and Schaal, S. (2008). Natural actor-critic. *Neurocomputing*, 71(7-9):1180–1190.

Polson, N. G., Sokolov, V., et al. (2017). Deep learning: a bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304.

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

Qian, X., Sailanbayev, A., Mishchenko, K., and Richtárik, P. (2019). Miso is making a comeback with better proofs and rates. *arXiv preprint arXiv:1906.01474*.

R Development Core Team (2008). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Razaviyayn, M., Hong, M., and Luo, Z.-Q. (2013). A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153.

Reddi, S., Zaheer, M., Sra, S., Poczos, B., Bach, F., Salakhutdinov, R., and Smola, A. (2018). A generic approach for escaping saddle points. In Storkey, A. and Perez-Cruz, F., editors, *Proceedings of the Twenty-First International Conference on Artificial Intelligence and Statistics*, volume 84 of *Proceedings of Machine Learning Research*, pages 1233–1242, Playa Blanca, Lanzarote, Canary Islands. PMLR.

Reddi, S. J., Hefny, A., Sra, S., Poczos, B., and Smola, A. (2016a). Stochastic variance reduction for nonconvex optimization. In *International conference on machine learning*, pages 314–323.

Reddi, S. J., Sra, S., Póczos, B., and Smola, A. (2016b). Fast incremental method for nonconvex optimization. *arXiv preprint arXiv:1603.06159*.

Rezende, D. J., Mohamed, S., and Wierstra, D. (2014). Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286.

Robbins, H. and Monro, S. (1951). A stochastic approximation method. *The Annals of Mathematical Statistics*, 22(3):400–407.

Robert, C. P. and Casella, G. (2010). *Metropolis–Hastings Algorithms*, pages 167–197. Springer New York, New York, NY.

Roberts, G. O., Gelman, A., and Gilks, W. R. (1997). Weak convergence and optimal scaling of random walk metropolis algorithms. *Ann. Appl. Probab.*, 7(1):110–120.

Roberts, G. O. and Rosenthal, J. S. (1997). Optimal scaling of discrete approximations to langevin diffusions. *J. R. Statist. Soc. B*, 60:255–268.

Roberts, G. O. and Rosenthal, J. S. (2011). Quantitative non-geometric convergence bounds for independence samplers. *Methodology and Computing in Applied Probability*, 13(2):391–403.

Roberts, G. O. and Tweedie, R. L. (1996). Exponential convergence of langevin distributions and their discrete approximations. *Bernoulli*, 2(4):341–363.

Roux, N. L., Schmidt, M., and Bach, F. R. (2012). A stochastic gradient method with an exponential convergence _rate for finite training sets. In Pereira, F., Burges, C. J. C., Bottou, L., and Weinberger, K. Q., editors, *Advances in Neural Information Processing Systems 25*, pages 2663–2671. Curran Associates, Inc.

Royer, C. W. and Wright, S. J. (2018). Complexity analysis of second-order line-search algorithms for smooth nonconvex optimization. *SIAM Journal on Optimization*, 28(2):1448–1477.

Rue, H., Martino, S., and Chopin, N. (2009). Approximate bayesian inference for latent gaussian models by using integrated nested laplace approximations. *Journal of the royal statistical society: Series b (statistical methodology)*, 71(2):319–392.

Samson, A., Lavielle, M., and Mentré, F. (2006). Extension of the saem algorithm to left-censored data in nonlinear mixed-effects model: Application to hiv dynamics model. *Computational Statistics & Data Analysis*, 51(3):1562–1574.

Savic, R. M., Mentré, F., and Lavielle, M. (2011). Implementation and evaluation of the SAEM algorithm for longitudinal ordered categorical data with an illustration in pharmacokinetics-pharmacodynamics. *The AAPS Journal*, 13(1):44–53.

Schmidt, M., Le Roux, N., and Bach, F. (2017). Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112.

Shapiro, A., Dentcheva, D., and Ruszczyński, A. (2009). *Lectures on stochastic programming: modeling and theory*. SIAM.

Snoeck, E., Chanu, P., Lavielle, M., Jacqmin, P., Jonsson, E., Jorga, K., Goggin, T., Grippo, J., Jumbe, N., and Frey, N. (2010). A comprehensive hepatitis c viral kinetic model explaining cure. *Clinical Pharmacology & Therapeutics*, 87(6):706–713.

Stan Development Team (2018). RStan: the R interface to Stan. R package version 2.17.3.

Stramer, O. and Tweedie, R. L. (1999). Langevin-type models i: Diffusions with given stationary distributions and their discretizations*. *Methodology And Computing In Applied Probability*, 1(3):283–306.

Sun, T., Sun, Y., and Yin, W. (2018). On Markov chain gradient descent. In Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Cesa-Bianchi, N., and Garnett, R., editors, *Advances in Neural Information Processing Systems 31*, pages 9918–9927. Curran Associates, Inc.

Sutskever, I., Martens, J., Dahl, G., and Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147.

Sutton, R. and Barto, A. (2018). *Reinforcement Learning: An Introduction, 2nd Edition*. MIT Press.

Sutton, R. S., McAllester, D. A., Singh, S. P., and Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. In *Advances in Neural Information Processing Systems*, pages 1057–1063.

Tadić, V. B. and Doucet, A. (2017). Asymptotic bias of stochastic gradient search. *The Annals of Applied Probability*, 27(6):3255–3304.

Thiesson, B., Meek, C., and Heckerman, D. (2001). Accelerating EM for large databases. *Machine Learning*, 45(3):279–299.

Titsias, M. K. and Papaspiliopoulos, O. (2018). Auxiliary gradient,Äźbased sampling algorithms. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 0(0).

Vanli, N. D., Gurbuzbalaban, M., and Ozdaglar, A. (2018). Global convergence rate of proximal incremental aggregated gradient methods. *SIAM Journal on Optimization*, 28(2):1282–1300.

Vapnik, V. (2013). *The nature of statistical learning theory*. Springer science & business media.

Verbeke, G. (1997). *Linear mixed models for longitudinal data*. Springer.

Vihola, M. (2012). Robust adaptive metropolis algorithm with coerced acceptance rate. *Statistics and Computing*, 22(5):997–1008.

Wainwright, M. J., Jordan, M. I., et al. (2008). Graphical models, exponential families, and variational inference. *Foundations and Trends® in Machine Learning*, 1(1–2):1–305.

Wang, Y. (2007). Derivation of various nonmem estimation methods. *Journal of Pharmacokinetics and pharmacodynamics*, 34(5):575–593.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015a). High dimensional EM algorithm: Statistical optimization and asymptotic normality. In Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., and Garnett, R., editors, *Advances in Neural Information Processing Systems 28*, pages 2521–2529. Curran Associates, Inc.

Wang, Z., Gu, Q., Ning, Y., and Liu, H. (2015b). High dimensional em algorithm: Statistical optimization and asymptotic normality. In *Advances in neural information processing systems*, pages 2521–2529.

Wei, G. C. G. and Tanner, M. A. (1990a). A monte carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Wei, G. C. G. and Tanner, M. A. (1990b). A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):699–704.

Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8(3-4):229–256.

WU, C. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, 11.

Wu, C. J. et al. (1983). On the convergence properties of the EM algorithm. *The Annals of statistics*, 11(1):95–103.

Xu, J., Hsu, D. J., and Maleki, A. (2016a). Global analysis of expectation maximization for mixtures of two gaussians. In Lee, D. D., Sugiyama, M., Luxburg, U. V., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems 29*, pages 2676–2684. Curran Associates, Inc.

Xu, J., Hsu, D. J., and Maleki, A. (2016b). Global analysis of Expectation Maximization for mixtures of two gaussians. In *Advances in Neural Information Processing Systems*, pages 2676–2684.

Xu, Y., Rong, J., and Yang, T. (2018). First-order stochastic algorithms for escaping from saddle points in almost linear time. In *Advances in Neural Information Processing Systems*, pages 5530–5540.

Zhang, Z. (2016). Parametric regression model for survival data: Weibull regression model as an example. *Ann Transl Med.*, 24.

List of Figures

List of Tables

université PARIS-SACLAY | École doctorale de mathématiques Hadamard (EDMH)

Titre : Optimisation Non Convexe pour Modèles à Données Latentes: Algorithmes, Analyse et Applications

Mots clés : approximation stochastique, optimisation non convexe, somme-finie, grande-echelle, données latentes, EM, MCMC, incremental, en ligne

Résumé : De nombreux problèmes en Apprentissage Statistique consistent à minimiser une fonction non convexe et non lisse définie sur un espace euclidien. Par exemple, les problèmes de maximisation de la vraisemblance et la minimisation du risque empirique en font partie. Les algorithmes d'optimisation utilisés pour résoudre ce genre de problèmes ont été largement étudié pour des fonctions convexes et grandement utilisés en pratique. Cependant, l'accrudescence du nombre d'observation dans l'évaluation de ce risque empirique ajoutée à l'utilisation de fonctions de perte de plus en plus sophistiquées représentent des obstacles. Ces obstacles requièrent d'améliorer les algorithmes existants avec des mis à jour moins coûteuses, idéalement indépendantes du nombre d'observations, et d'en garantir le comportement théorique sous des hypothèses moins restrictives, telles que la non convexité de la fonction à optimiser.

Dans ce manuscrit de thèse, nous nous intéressons à la minimisation de fonctions objectives pour des modèles à données latentes, *i.e.,* lorsque les données sont partiellement observées ce qui inclut le sens conventionnel des données manquantes mais est un terme plus général que cela. Dans une première partie, nous considérons la minimisation d'une fonction (possiblement) non convexe et non lisse en utilisant des mises à jour *incrémentales* et *en ligne*. Nous proposons et analysons plusieurs algorithmes à travers quelques applications. Dans une seconde partie, nous nous concentrons sur le problème de maximisation de vraisemblance non convexe en ayant recourt à l'algorithme EM et ses variantes stochastiques. Nous en analysons plusieurs versions rapides et moins coûteuses et nous proposons deux nouveaux algorithmes du type EM dans le but d'accélérer la convergence des paramètres estimés.

Title : Non-Convex Optimization for Latent Data Models: Algorithms, Analysis and Applications

Keywords: stochastic approximation, non-convex optimization, finite-sum, large-scale, latent data, EM, MCMC, incremental, online

Abstract: Many problems in machine learning pertain to tackling the minimization of a possibly non-convex and non-smooth function defined on a Euclidean space. Examples include topic models, neural networks or sparse logistic regression. Optimization methods, used to solve those problems, have been widely studied in the literature for convex objective functions and are extensively used in practice. However, recent breakthroughs in statistical modeling, such as deep learning, coupled with an explosion of data samples, require improvements of non-convex optimization procedure for large datasets. This thesis is an attempt to address those two challenges by developing algorithms with cheaper updates, ideally independent of the number of samples, and improving the theoretical understanding of non-convex optimization that remains rather limited. In this manuscript, we are interested in the minimization of such objective functions for latent data models, *i.e.,* when the data is partially observed which includes the conventional sense of missing data but is much broader than that. In the first part, we consider the minimization of a (possibly) non-convex and non-smooth objective function using *incremental* and *online* updates. To that end, we propose several algorithms exploiting the latent structure to efficiently optimize the objective and illustrate our findings with numerous applications. In the second part, we focus on the maximization of non-convex likelihood using the EM algorithm and its stochastic variants. We analyze several faster and cheaper algorithms and propose two new variants aiming at speeding the convergence of the estimated parameters.

Université Paris-Saclay
Espace Technologique / Immeuble Discovery
Route de l'Orme aux Merisiers RD 128 / 91190 Saint-Aubin, France