HAL
archives-ouvertes.fr

Stochastic Proximal Gradient Algorithms for Penalized Mixed Models

Gersende Fort, Edouard Ollier, Adeline Samson

To cite this version:

Gersende Fort, Edouard Ollier, Adeline Samson. Stochastic Proximal Gradient Algorithms for Penalized Mixed Models. Statistics and Computing, Springer Verlag (Germany), 2019, 29 (2), pp.231-253. 10.1007/s11222-018-9805-7 . hal-01526281

HAL Id: hal-01526281

https://hal.archives-ouvertes.fr/hal-01526281

Submitted on 22 May 2017

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

arXiv:1704.08891v1 [stat.CO] 28 Apr 2017

Stochastic Proximal Gradient Algorithms for Penalized Mixed Models

Gersende Fort[1], Edouard Ollier[2,3], and Adeline Samson[4]

[1]IMT UMR5219, Université de Toulouse, CNRS; F-31062 Toulouse Cedex 9, France.
[2]INSERM, U1059, Dysfonction Vasculaire et Hémostase, Saint Etienne, France.
[3]U.M.P.A., Ecole Normale Supérieure de Lyon, CNRS UMR 5669; INRIA, Project-team NUMED. 46 Allée d'Italie, 69364 Lyon Cedex 07, France
[4]Université Grenoble-Alpes, Laboratoire Jean Kuntzmann, UMR CNRS 5224

May 1, 2017

Abstract

Motivated by penalized likelihood maximization in complex models, we study optimization problems where neither the function to optimize nor its gradient have an explicit expression, but its gradient can be approximated by a Monte Carlo technique. We propose a new algorithm based on a stochastic approximation of the Proximal-Gradient (PG) algorithm. This new algorithm, named Stochastic Approximation PG (SAPG) is the combination of a stochastic gradient descent step which - roughly speaking - computes a smoothed approximation of the past gradient along the iterations, and a proximal step. The choice of the step size and the Monte Carlo batch size for the stochastic gradient descent step in SAPG are discussed. Our convergence results cover the cases of biased and unbiased Monte Carlo approximations. While the convergence analysis of the Monte Carlo-PG is already addressed in the literature (see Atchadé et al. [2016]), the convergence analysis of SAPG is new. The two algorithms are compared on a linear mixed effect model as a toy example. A more challenging application is proposed on non-linear mixed effect models in high dimension with a pharmacokinetic data set including genomic covariates. To our best knowledge, our work provides the first convergence result of a numerical method designed to solve penalized Maximum Likelihood in a non-linear mixed effect model.

Keywords: Proximal-Gradient algorithm; Stochastic Gradient; Stochastic EM algorithm; Stochastic Approximation; Non-linear mixed effect models.

1 Introduction

Many problems in computational statistics reduce to the maximization of a criterion

$$\mathrm{argmax}_{\theta \in \mathbb{R}^d} F(\theta), \qquad \text{where } F := \ell - g, \tag{1}$$

and the functions ℓ, g satisfy

H1. *the function $g : \mathbb{R}^d \to [0, +\infty]$ is convex, not identically $+\infty$, and lower semi-continuous.*

H 2. *the function* $\ell : \mathbb{R}^d \to \mathbb{R} \cup \{-\infty\}$ *is continuously differentiable on* $\Theta := \{\theta \in \mathbb{R}^d : g(\theta) + |\ell(\theta)| < \infty\}$ *and its gradient is of the form*

$$\nabla \ell(\theta) = \nabla \phi(\theta) + \Psi(\theta)\overline{S}(\theta), \quad \text{with } \overline{S}(\theta) := \int_{\mathsf{Z}} S(z)\pi_\theta(z)\mu(\mathrm{d}z); \tag{2}$$

∇ *denotes the gradient operator. The measurable functions* $\nabla\phi : \mathbb{R}^d \to \mathbb{R}^d$ *and* $\Psi : \mathbb{R}^d \to \mathbb{R}^{d\times q}$ *are known but the expectation* $\overline{S}$ *of the function* $S : \mathsf{Z} \to \mathbb{R}^q$ *with respect to some probability distribution* $\pi_\theta \mathrm{d}\mu$ *on a measurable subset* $(\mathsf{Z}, \mathcal{Z})$ *of* $\mathbb{R}^p$ *may be intractable. Furthermore, there exists a finite non-negative constant* L *such that for all* $\theta, \theta' \in \Theta$,

$$\|\nabla \ell(\theta) - \nabla \ell(\theta')\| \leq L\|\theta - \theta'\|; \tag{3}$$

$\|\cdot\|$ *is the Euclidean norm.*

Examples of functions ℓ satisfying (2) are given below. We are interested in numerical methods for solving (1), robust to the case when neither ℓ nor its gradient have an explicit expression.

Such an optimization problem occurs for example when computing a penalized maximum likelihood estimator in some parametric model indexed by $\theta \in \mathbb{R}^d$: ℓ denotes the log-likelihood of the observations Y (the dependence upon Y is omitted) and g is the penalty term.
The optimization problem (1) covers the computation of the maximum when the parameter θ is restricted to a closed convex subset Θ of $\mathbb{R}^d$; in that case, g is the characteristic function of Θ i.e. $g(\theta) = 0$ for any $\theta \in \Theta$ and $g(\theta) = +\infty$ otherwise. It also covers the case when g is the ridge, the lasso or the elastic net penalty; and more generally, the case when g is the sum of lower semi-continuous non-negative convex functions.
A first example of such a function ℓ is given by the log-likelihood in a latent variable model with complete likelihood from the q-parameter exponential family (see e.g. Bickel and Doksum [2015] and Bartholomew et al. [2011] and the references therein). In that case, ℓ is of the form

$$\theta \mapsto \ell(\theta) := \log \int_{\mathsf{Z}} \exp\left(\phi(\theta) + \langle S(z), \psi(\theta)\rangle\right) \mu(\mathrm{d}z), \tag{4}$$

where $\langle a, b\rangle$ denotes the scalar of two vectors $a, b \in \mathbb{R}^l$, $\phi : \mathbb{R}^d \to \mathbb{R}$, $\psi : \mathbb{R}^d \to \mathbb{R}^q$ and $S : \mathsf{Z} \to \mathbb{R}^q$ are measurable functions, and μ is a σ-finite positive measure on $\mathcal{Z}$. The quantity $\theta \mapsto \phi(\theta) + \langle S(Z), \psi(\theta)\rangle$ is known as the complete log-likelihood, and Z is the latent data vector. Under regularity conditions, we have

$$\nabla \ell(\theta) = \nabla\phi(\theta) + \mathrm{J}\,\psi(\theta) \int_{\mathsf{Z}} S(z)\ \pi_\theta(z)\mu(\mathrm{d}z), \quad \text{with } \pi_\theta(z) := \frac{\exp(\langle S(z), \psi(\theta)\rangle)}{\int_{\mathsf{Z}} \exp(\langle S(u), \psi(\theta)\rangle)\mu(\mathrm{d}u)}, \tag{5}$$

where $\mathrm{J}\,\psi(\theta)$ denotes the transpose of the jacobian matrix of the function ψ at θ.
A second example is given by the log-likelihood of N independent observations $(\mathsf{Y}_1, \cdots, \mathsf{Y}_N)$ from a log-linear model for Markov random fields. In this model, ℓ is given by

$$\theta \mapsto \ell(\theta) := \sum_{k=1}^{N} \langle S(\mathsf{Y}_k), \theta\rangle - N \log \int_{\mathsf{Z}} \exp\left(\langle S(z), \theta\rangle\right) \mu(\mathrm{d}z). \tag{6}$$

The function $\theta \mapsto \int_{\mathsf{Z}} \exp\left(\langle S(z), \theta\rangle\right) \mu(\mathrm{d}z)$ is known as the partition function. Under regularity conditions, we have

$$\nabla \ell(\theta) = \sum_{k=1}^{N} S(\mathsf{Y}_k) - N \int_{\mathsf{Z}} S(z)\ \pi_\theta(z) \mu(\mathrm{d}z), \quad \text{with}\, \pi_\theta(z) := \frac{\exp\left(\langle S(z), \theta\rangle\right)}{\int \exp\left(\langle S(u), \theta\rangle\right)\ \mu(\mathrm{d}u)}. \tag{7}$$

In these two examples, the integrals in Eqs. (4) to (7) are intractable except for toy examples: neither the function ℓ nor its gradient are available. Nevertheless, all the integrals in Eqs. (4)-(7) can be approximated by a Monte Carlo sum (see e.g. Robert and Casella [2004]). In the first example, this Monte Carlo approximation consists in imputing the missing variables z; it is known that such an imputation is far more efficient when the Monte Carlo samples are drawn under $\pi_\theta \mathrm{d}\mu$, i.e. the a posteriori distribution of the missing variables given the observations (see Eq. (5)) than when they are drawn under the a priori distribution. This remark is the essence of the Expectation Maximization (EM) algorithm (Dempster et al. [1977]), a popular iterative procedure for maximizing the log-likelihood ℓ in latent variable models.

In this paper, we are interested in first order optimization methods to solve (1), that is methods based on the gradient. In Section 2.1, we propose two stochastic first-order descent methods, which are stochastic perturbations of the Proximal-Gradient (PG) algorithm (see e.g. Beck and Teboulle [2009], Combettes and Pesquet [2011], Parikh and Boyd [2013] for literature reviews on Proximal-Gradient algorithms). The two algorithms are the Monte Carlo Proximal-Gradient algorithm (MCPG) and the Stochastic Approximation Proximal-Gradient algorithm (SAPG), which differ in the approximation of the gradient $\nabla \ell$ and more precisely, of the intractable integral $\overline{S}(\theta)$ (see Eq. (2)). In MCPG, at each iteration n of the algorithm, this expectation evaluated at the current point θ_n is approximated by a Monte Carlo sum computed from samples $\{Z_{1,n}, \cdots, Z_{m_{n+1},n}\}$ approximating $\pi_{\theta_n} \mathrm{d}\mu$. In SAPG, the approximation is computed as a Monte Carlo sum based on all the points drawn during all the previous iterations of the algorithm $\{Z_{i,j}, i \leq m_j, j \leq n+1\}$.

When ℓ is the log-likelihood of a latent variable model, we prove in Section 2.2 that our algorithms are Generalized EM algorithms (see e.g. McLachlan and Krishnan [2008], Ng et al. [2012]) combined with a stochastic E-step: in MCPG and SAPG, the stochastic E-step mimics the E-step of the Monte Carlo EM (Wei and Tanner [1990], Levine and Fan [2004]) and the E-step of the Stochastic Approximation EM (Delyon et al. [1999]), respectively.

MCPG and SAPG are compared through a toy example and the choice of some design parameters is discussed in Section 2.3. A more challenging application to penalized inference in a mixed effect model is detailed in Section 3. Mixed models are applied to analyze repeated data in a population of subjects. The inter-subject variability is modeled by considering individual parameters of the regression function as random variables. The N independent vectors of observations $(\mathsf{Y}_k, k = 1, \ldots, N)$ of the N subjects are modeled by

$$\mathsf{Y}_k = f(t_k, Z^{(k)}) + \varepsilon_k, \tag{8}$$

with individual latent variable $Z^{(k)}$ independent of the measurement error vector ε_k and f the regression function that depends on the vector of observation times t_k. Mixed models thus enter the class of models (4) with latent variables $Z = (Z^{(1)}, \ldots, Z^{(N)})$. These models are widely used in pharmacokinetic (PK) and pharmacodynamic. PK aims at modeling the evolution along time of a drug concentration in the body to understand the influence of covariates. Covariates can be gender, age, weight, but also levels of genes expression. The statistical issue is thus the selection of the most influent covariates in high dimension for a latent variable model. The signal is typically sparse, meaning that the number of influent covariates is expected to be small. The selection problem can therefore be treated through the optimization of a penalized

version of the log-likelihood (4). In non-linear mixed models, the optimization problem is not explicit and tentatives have been proposed with stochastic penalized versions of the EM [Bertrand and Balding, 2013, Ollier et al., 2016, Chen et al., 2017]. To our best knowledge, stochastic proximal gradient algorithms have not been proposed for mixed models.

Finally, Section 4 is devoted to the convergence analysis of MCPG and SAPG. These algorithms can be seen as perturbed proximal gradient algorithms when the perturbation comes from replacing the exact quantity $\overline{S}(\theta_n)$ by a Monte Carlo approximation S_{n+1} at each iteration of the algorithm. Our convergence analysis covers the case when the points $\{Z_{1,n}, \cdots, Z_{m_{n+1},n}\}$ are sampled from a Markov chain Monte Carlo sampler (MCMC) with target distribution $\pi_{\theta_n} \mathrm{d}\mu$ - and therefore, it also covers the case of i.i.d. draws. This implies that the estimator S_{n+1} of $\overline{S}(\theta_n)$ may be biased. There exist many contributions in the literature on the convergence of perturbed proximal-gradient algorithms when ℓ is concave, but except in the works by Atchadé et al. [2016] and Combettes and Pesquet [2015], most of them assume that the error $S_{n+1} - \overline{S}(\theta_n)$ is unbiased and gets small when $n \to \infty$ (see e.g. Rosasco et al. [2014], Combettes and Pesquet [2016], Rosasco et al. [2016], Lin et al. [2015]). In this paper, we provide sufficient conditions for the almost-sure convergence of MCPG and SAPG under the assumption that ℓ is concave and with no assumptions on the bias of $S_{n+1} - \overline{S}(\theta_n)$. The convergence analysis of MCPG is a simple adaptation of [Atchadé et al., 2016, Section 4]; to our best knowledge, the convergence of SAPG is a new result.

2 Stochastic Proximal-Gradient based algorithms

In this section, we provide first-order based algorithms for solving (1) under the assumptions H1 and H2, when the expectation $\overline{S}(\theta)$ in (2) is intractable.

2.1 The MCPG and SAPG algorithms

We propose two iterative algorithms, each update relies on the combination of a gradient step and a proximal operator.

The proximal map (Moreau [1962], see also Bauschke and Combettes [2011], Parikh and Boyd [2013]) associated to a convex function g is defined for any $\gamma > 0$ and $\theta \in \mathbb{R}^d$ by

$$\mathrm{Prox}_{\gamma,g}(\theta) := \mathrm{argmin}_{\tau \in \Theta} \left\{ g(\tau) + \frac{1}{2\gamma} \|\theta - \tau\|^2 \right\}. \tag{9}$$

Note that under H1, for any $\gamma > 0$ and $\theta \in \mathbb{R}^d$, there exists an unique point τ minimizing the RHS of (9). This proximal operator may have an explicit expression. For example, when g is the characteristic function

$$g(\theta) := \begin{cases} 0 & \text{if } \theta \in \Theta \\ +\infty & \text{otherwise,} \end{cases}$$

for some closed convex set $\Theta \subseteq \mathbb{R}^d$, then $g(\theta)$ is the projection of θ on Θ. Another example is the elastic net penalty i.e. $g_{\lambda,\alpha}(\theta) := \lambda \left(\frac{1-\alpha}{2} \sum_{i=1}^d \theta_i^2 + \alpha \sum_{i=1}^d |\theta_i| \right)$ with $\theta = (\theta_1, \cdots, \theta_d)$, $\lambda > 0$ and $\alpha \in (0, 1]$, then for any component $i \in \{1, \cdots, d\}$,

$$(\mathrm{Prox}_{\gamma,g}(\theta))_i = \frac{1}{1 + \gamma\lambda(1-\alpha)} \begin{cases} 0 & \text{if } |\theta_i| \leq \gamma\lambda\alpha, \\ \theta_i - \gamma\lambda\alpha & \text{if } \theta_i \geq \gamma\lambda\alpha, \\ \theta_i + \gamma\lambda\alpha & \text{if } \theta_i \leq -\gamma\lambda\alpha. \end{cases}$$

The proximal gradient algorithm for solving the optimization problem (1) produces a sequence $\{\theta_n, n \geq 0\}$ as follows: given a $(0, 1/L]$-valued sequence $\{\gamma_n, n \geq 0\}$,

$$\begin{aligned}\theta_{n+1} &= \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n + \gamma_{n+1}\nabla\ell(\theta_n)\right) \\ &= \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n + \gamma_{n+1}\{\nabla\phi(\theta_n) + \Psi(\theta_n)\overline{S}(\theta_n)\}\right).\end{aligned} \tag{10}$$

This update scheme can be explained as follows: by H2, we have for any $L \leq \gamma_{n+1}^{-1}$,

$$\begin{aligned}&F(\theta) = \ell(\theta) - g(\theta) \\ &\geq \ell(\theta_n) - \langle\nabla\ell(\theta_n), \theta - \theta_n\rangle - \frac{1}{2\gamma_{n+1}}\|\theta - \theta_n\|^2 - g(\theta).\end{aligned}$$

This minorizing function is equal to $F(\theta_n)$ at the point θ_n; the maximization (w.r.t. θ) of the RHS yields θ_{n+1} given by (10). The proximal-gradient is therefore a Minorize - Majorization (MM) algorithm and the ascent property holds: $F(\theta_{n+1}) \geq F(\theta_n)$ for all n. Sufficient conditions for the convergence of the proximal-gradient algorithm (10) can be derived from convergence analysis of MM algorithms (see e.g. Zangwill [1969], Meyer [1976]), or of Proximal-Gradient algorithms (see e.g. Combettes and Wajs [2005], Parikh and Boyd [2013]).

In the case $\overline{S}(\theta)$ can not be computed, we propose two strategies for a Monte Carlo approximation. At iteration $n+1$, given the current value of the parameter θ_n, m_{n+1} points $\{Z_{1,n}, \cdots, Z_{m_{n+1},n}\}$ from the path of a Markov chain with target distribution $\pi_{\theta_n}\mathrm{d}\mu$ are sampled. A first strategy consists in replacing $\overline{S}(\theta_n)$ by a Monte Carlo mean:

$$S_{n+1}^{\mathrm{mc}} := \frac{1}{m_{n+1}}\sum_{j=1}^{m_{n+1}} S(Z_{j,n}). \tag{11}$$

A second strategy, inspired by stochastic approximation methods (see e.g. Benveniste et al. [1990], Kushner and Yin [2003]) consists in replacing $\overline{S}(\theta_n)$ by a stochastic approximation

$$S_{n+1}^{\mathrm{sa}} := (1 - \delta_{n+1})S_n^{\mathrm{sa}} + \frac{\delta_{n+1}}{m_{n+1}}\sum_{j=1}^{m_{n+1}} S(Z_{j,n}), \tag{12}$$

where $\{\delta_n, n \geq 0\}$ is a deterministic $[0, 1]$-valued sequence. These two strategies yield respectively the Monte Carlo Proximal-Gradient (MCPG) algorithm (see Algorithm 1) and the Stochastic Approximation Proximal-Gradient (SAPG) algorithm (see Algorithm 2).

input : The initial values $\theta_0 \in \Theta$ and $Z_{m_0,-1} := z_\star$, a $(0, 1/L]$-valued sequence $\{\gamma_n, n \geq 0\}$ and an integer valued sequence $\{m_n, n \geq 0\}$
output: The sequence $\{\theta_n, n \geq 0\}$
for $n \geq 0$ **do**
  Simulation-step ;
  sample a path $Z_{1,n}, \cdots, Z_{m_{n+1},n}$ of a Markov chain with invariant distribution $\pi_{\theta_n}\mathrm{d}\mu$ and started from $Z_{m_n,n-1}$;
  Expectation step ;
  Compute S_{n+1}^{mc} as in (11) ;
  Maximization step ;
  Set $\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n + \gamma_{n+1}\{\nabla\phi(\theta_n) + \Psi(\theta_n)S_{n+1}^{\mathrm{mc}}\}\right)$

Algorithm 1: The Monte Carlo Proximal-Gradient algorithm for the maximization of $\ell - g$

input : The initial values $\theta_0 \in \Theta$, $Z_{m_0,-1} := z_\star$ and $S_0^{\mathrm{sa}} := s_\star$, a $(0, 1/L]$-valued sequence $\{\gamma_n, n \geq 0\}$, a $[0,1]$-valued sequence $\{\delta_n, n \geq 0\}$ and an integer valued sequence $\{m_n, n \geq 0\}$

output: The sequence $\{\theta_n, n \geq 0\}$

for $n \geq 0$ **do**

  Simulation-step ;

  sample a path $Z_{1,n}, \cdots, Z_{m_{n+1},n}$ of a Markov chain with invariant distribution $\pi_{\theta_n} \mathrm{d}\mu$ and started from $Z_{m_n,n-1}$;

  Expectation step ;

  Compute S_{n+1}^{sa} as in (12) ;

  Maximization step ;

  Set $\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1},g}\left(\theta_n + \gamma_{n+1}\{\nabla\phi(\theta_n) + \Psi(\theta_n) S_{n+1}^{\mathrm{sa}}\}\right)$

Algorithm 2: The Stochastic Approximation Proximal-Gradient algorithm for the maximization of $\ell - g$

In Section 4, we prove the convergence of MCPG to the maxima of F when ℓ is concave, for different choices of the sequences $\{\gamma_n, m_n, n \geq 0\}$ including decreasing or constant stepsizes $\{\gamma_n, n \geq 0\}$ and respectively, constant or increasing batch size $\{m_n, n \geq 0\}$. We also establish the convergence of SAPG to the maxima (in the concave case); only the case of constant batch size $\{m_n, n \geq 0\}$ and decreasing stepsize $\{\gamma_n, n \geq 0\}$ is studied, since this framework corresponds to the *stochastic approximation* one from which the update rule (12) is inherited (see Delyon et al. [1999]). From a numerical point of view, the choice of the sequences $\{\gamma_n, n \geq 0\}$, $\{\delta_n, n \geq 0\}$ and $\{m_n, n \geq 0\}$ is discussed on a toy example in Section 2.3.

2.2 Case of latent variable models from the exponential family

In this section, we consider the case when ℓ is given by (4). A classical approach to solve penalized maximum likelihood problems in latent variables models with complete likelihood from the exponential family is the Expectation-Maximization (EM) algorithm or a generalization called the Generalized EM (GEM) algorithm (Dempster et al. [1977], McLachlan and Krishnan [2008], Ng et al. [2012]). The goal here is to show that MCPG and SAPG are stochastic perturbations of a GEM algorithm.

The EM algorithm is an iterative algorithm: at each iteration, given the current parameter θ_n, the quantity $\mathcal{Q}(\theta|\theta_n)$, defined as the conditional expectation of the complete log-likelihood under the a posteriori distribution for the current fit of the parameters, is computed:

$$\mathcal{Q}(\theta|\theta') := \phi(\theta) + \left\langle \overline{S}(\theta'), \psi(\theta) \right\rangle . \tag{13}$$

The EM sequence $\{\theta_n, n \geq 0\}$ for the maximization of the penalized log-likelihood $\ell - g$ is given by (see [McLachlan and Krishnan, 2008, Section 1.6.1.])

$$\theta_{n+1} = \mathrm{argmax}_{\theta \in \Theta} \left\{ \phi(\theta) + \left\langle \overline{S}(\theta_n), \psi(\theta) \right\rangle - g(\theta) \right\} . \tag{14}$$

When $\overline{S}(\theta)$ is intractable, it was proposed to replace $\overline{S}(\theta_n)$ in this EM-penalized algorithm by an approximation S_{n+1} - see Algorithm 3. When $S_{n+1} = S_{n+1}^{mc}$ (see (11)), this yields the so-called Monte Carlo-EM penalized algorithm (MCEM-pen), adapted from the MCEM by Wei and Tanner [1990], Levine and Fan [2004]. Another popular strategy is to replace $\overline{S}(\theta_n)$ by S_{n+1}^{sa} (see (12)) yielding to the so-called Stochastic Approximation-EM penalized algorithm (SAEM-pen) - [see Delyon et al., 1999, for the unpenalized version].

input : The initial value $\theta_0 \in \Theta$
output: The sequence $\{\theta_n, n \geq 0\}$
for $n \geq 0$ **do**
  E-step ;
  Compute an approximation S_{n+1} of $\overline{S}(\theta_n)$;
  M-step ;
  Set $\theta_{n+1} = \mathrm{argmax}_\theta \{\phi(\theta) + \langle S_{n+1}, \psi(\theta)\rangle - g(\theta)\}$

Algorithm 3: Perturbed EM-penalized algorithms for the maximization of $\ell - g$

When the maximization of (13) is not explicit, the update of the parameter is modified as follows:

$$\begin{aligned} \theta_{n+1} \text{ s.t. } & \phi(\theta_{n+1}) + \left\langle \overline{S}(\theta_n), \psi(\theta_{n+1}) \right\rangle - g(\theta_{n+1}) \\ & \geq \phi(\theta_n) + \left\langle \overline{S}(\theta_n), \psi(\theta_n) \right\rangle - g(\theta_n); \end{aligned} \tag{15}$$

this yields the Generalized EM-penalized algorithm (GEM-pen). This update rule still produces a sequence $\{\theta_n, n \geq 0\}$ satisfying the ascent property $F(\theta_{n+1}) \geq F(\theta_n)$ which is the key property for the convergence of EM (Wu [1983]). Here again, the approximations (11) and (12) can be plugged in the GEM-pen update (15) when $\overline{S}$ is not explicit.

We show in the following proposition that the sequence $\{\theta_n, n \geq 0\}$ produced by the Proximal-Gradient algorithm (10) is a GEM-pen sequence since it satisfies the inequality (15). As a consequence, MCPG and SAPG are stochastic GEM-pen algorithms.

Proposition 1. *Let g satisfying H1 and ℓ be of the form* (4) *with continuously differentiable functions $\phi : \mathbb{R}^d \to \mathbb{R}$, $\psi : \mathbb{R}^d \to \mathbb{R}^q$ and $S : \mathsf{Z} \to \mathbb{R}^q$. Set $\Theta := \{g + |\ell| < \infty\}$. Define $\overline{S} : \Theta \to \mathbb{R}^q$ by $\overline{S}(\theta) := \int_{\mathsf{Z}} S(z)\, \pi_\theta(z)\, \mu(\mathrm{d}z)$ where π_θ is given by* (5)*. Assume that there exists a constant $L > 0$ such that for any $s \in \overline{S}(\Theta)$, and any $\theta, \theta' \in \Theta$,*

$$\|\nabla\phi(\theta) - \nabla\phi(\theta') + (\mathrm{J}\,\psi(\theta) - \mathrm{J}\,\psi(\theta'))\, s\| \leq L\|\theta - \theta'\|.$$

Let $\{\gamma_n, n \geq 0\}$ be a (deterministic) positive sequence such that $\gamma_n \in (0, 1/L]$ for all $n \geq 0$.
Then the proximal-gradient algorithm (10) *is a GEM-pen algorithm for the maximization of $\ell - g$.*

The proof is postponed in Section 5.1.

2.3 Comparison on a toy example

We compare MCPG, SAPG, SAEM-pen (i.e. Algorithm 3 applied with $S_{n+1} = S^{\mathrm{sa}}_{n+1}$) and EM-pen for the maximization of a penalized log-likelihood in a toy latent variable model: a linear mixed model. In this example, the exact algorithm EM-pen (see (14)) applies: the quantity $\overline{S}(\theta)$ is an explicit expectation under a Gaussian distribution π_θ. Therefore, we use this example *(i)* to illustrate the convergence of the three stochastic methods to the same limit point as EM-pen, *(ii)* to compare the two approximations S^{mc}_{n+1} and S^{sa}_{n+1} of $\overline{S}(\theta_n)$ in a GEM-pen approach, and *(iii)* to study the effect of relaxing the M-step by comparing the GEM-pen and EM-pen approaches namely SAPG and SAEM-pen. The objective is also to understand the influence of the step size sequences $\{\gamma_n, \delta_n, n \geq 1\}$.

The example is a mixed model (8), where the function f is linear in the latent variable Z. More precisely, we observe data for N independent subjects, each individual data being a vector of size J. For the subject k, $k = 1, \cdots, N$, Y_{kj} is the j-th measure at time t_{kj}, $j = 1, \cdots, J$. Set $\mathsf{Y}_k := (\mathsf{Y}_{k1}, \ldots, \mathsf{Y}_{kJ})$. It is assumed

that $\{\mathsf{Y}_k, k = 1, \cdots, N\}$ are independent and for all $k = 1, \cdots, N$,

$$\begin{aligned} &Y_{kj}|Z^{(k)} \overset{ind}{\sim} \mathcal{N}\left(\left\langle Z^{(k)}, \bar{t}_{kj}\right\rangle, 1\right), \\ &\bar{t}_{kj} := \begin{bmatrix} 1 \\ t_{kj} \end{bmatrix} \qquad j = 1, \cdots, J; \end{aligned} \tag{16}$$

that is, a linear regression model with individual random intercept and slope, the 2-dimensional vector being denoted $Z^{(k)}$. The latent variable is $Z = (Z^{(1)}, \ldots, Z^{(N)})$. Furthermore,

$$Z^{(k)} \overset{ind}{\sim} \mathcal{N}_2(X_k\theta, I_2); \tag{17}$$

here, $\theta \in \mathbb{R}^{2(D+1)}$ is an unknown parameter and the design matrix $X_k \in \mathbb{R}^{2\times 2(D+1)}$ is known

$$X_k := \begin{bmatrix} 1 & X_{k1} & \ldots & X_{kD} & 0 & 0 & \ldots & 0 \\ 0 & 0 & \ldots & 0 & 1 & X_{k1} & \ldots & X_{kD} \end{bmatrix}. \tag{18}$$

The optimization problem (1) that we consider is the log-likelihood $\ell(\theta)$ penalized by a lasso penalty: the objective is the selection of the influent covariates

$$(X_{k1}, \ldots, X_{kD})$$

on the two components of $Z^{(k)}$. We thus penalize all the elements except θ_1 and θ_{D+2} which correspond to the two intercepts; hence, we set

$$g(\theta) := \lambda \sum_{r \neq \{1, D+2\}} |\theta_r|.$$

The above model is a latent variable model with complete log-likelihood equal to - up to an additive constant

$$-\frac{1}{2}\sum_{k=1}^{N}\left\{\sum_{j=1}^{J}\left(\mathsf{Y}_{kj} - \left\langle Z^{(k)}, \bar{t}_{kj}\right\rangle\right)^2 + \left\|Z^{(k)} - X_k\theta\right\|^2\right\}.$$

It is of the form $\phi(\theta) + \langle S(z), \psi(\theta)\rangle$ by setting (with $(\cdot)'$ denoting the transpose of a matrix)

$$\begin{aligned} &\phi(\theta) := -\frac{1}{2}\theta'\left(\sum_{k=1}^{N} X_k' X_k\right)\theta - \frac{1}{2}\sum_{k=1}^{N}\sum_{j=1}^{J}\mathsf{Y}_{kj}^2, \\ &\psi(\theta) := \begin{bmatrix} 1 \\ \theta \end{bmatrix} \in \mathbb{R}^{1+2(D+1)}, \\ &S(z^{(1)}, \cdots, z^{(N)}) := \\ &-\frac{1}{2}\sum_{k=1}^{N}\begin{bmatrix} z^{(k)\prime}(I + T_k)z^{(k)} - 2\left\langle z^{(k)}, \bar{\mathsf{Y}}_k\right\rangle \\ -2X_k' z^{(k)} \end{bmatrix}, \\ &T_k := \sum_{j=1}^{J}\bar{t}_{kj}\bar{t}_{kj}', \\ &\bar{\mathsf{Y}}_k := \sum_{j=1}^{J}\mathsf{Y}_{kj}\bar{t}_{kj}. \end{aligned}$$

The a posteriori distribution π_θ is a Gaussian distribution on $\mathbb{R}^{2N}$, equal to the product of N Gaussian distributions on $\mathbb{R}^2$:

$$\pi_\theta(z^{(1)}, \cdots, z^{(N)}) := \prod_{k=1}^{N} \mathcal{N}_2\left((I+T_k)^{-1}\left(\bar{\mathsf{Y}}_k + X_k\theta\right), (I+T_k)^{-1}\right)[z^{(k)}]. \tag{19}$$

Hence, $\overline{S}(\theta)$ is explicit and given by

$$\overline{S}(\theta) = -\frac{1}{2}\sum_{k=1}^{N}\begin{bmatrix}\mathrm{Trace}((I+T_k)\Sigma_k) - 2\bar{\mathsf{Y}}_k'(I+T_k)^{-1}\left(\bar{\mathsf{Y}}_k + X_k\theta\right) \\ -2X_k'(I+T_k)^{-1}\left(\bar{\mathsf{Y}}_k + X_k\theta\right)\end{bmatrix} \tag{20}$$

with

$$\Sigma_k := (I+T_k)^{-1} + (I+T_k)^{-1}\left(\bar{\mathsf{Y}}_k + X_k\theta\right)\left(\bar{\mathsf{Y}}_k + X_k\theta\right)'(I+T_k)^{-1}. \tag{21}$$

Finally, note that in this example, the function ℓ is explicit and given by (up to an additive constant)

$$\ell(\theta) = -\frac{1}{2}\theta'\left(\sum_{k=1}^{N} X_k'X_k\right)\theta + \frac{1}{2}\sum_{k=1}^{N}(\bar{\mathsf{Y}}_k + X_k\theta)'(I+T_k)^{-1}(\bar{\mathsf{Y}}_k + X_k\theta).$$

Thus ℓ is a concave function.
A data set is simulated using this model with $N = 40$, $J = 8$, $D = 300$ and $t_{kj} \in \{0.25, 4, 6, 8, 10, 12, 14, 16\}$, $\forall k \in \{1, ..., N\}$. The design components $(X_{k1}, \ldots, X_{kD})$ (see eq (18)) are drawn from $\mathcal{N}(0, \Gamma)$ with $\Gamma_{rr'} = 0.5^{|r-r'|}$ $(r, r' = 1, ..., 300)$. To sample the observations, we use a parameter vector $\theta^\star$ defined as follows: $\theta_1^\star = \theta_{D+2}^\star = 1$; the other components are set to zero, except 12 components randomly selected (6 among the components $\{2, \cdots, D+1\}$ and 6 among the components $\{D+3, \cdots, 2D+2\}$) and chosen uniformly in $[0.5, 1.5]$ - see the last row on Figure 4.
The sequences $\{\gamma_n, n \geq 0\}$ and $\{\delta_n, n \geq 0\}$ are defined as follows: $\gamma_n = 0.004$ and $\delta_n = 0.5$ for $n \leq 200$, then when $n > 200$, $\gamma_n = 0.004n^{-\alpha}$ and $\delta_n = 0.5n^{-\beta}$; three different pairs (α, β) are considered: $\alpha = 0.9$ and $\beta = 0.4$, $\alpha = 0.6$ and $\beta = 0.1$, $\alpha = 0.5$ and $\beta = 0.5$. The algorithms are implemented with a fixed batch size $m_n = 60$. 100 independent runs of each algorithm are performed. For the penalty term, we set $\lambda = 50$. In MCPG, SAPG and SAEM-pen, the simulation step at iteration $(n+1)$ relies on exact sampling from π_{θ_n} - see (19); therefore, in this toy example, the Monte Carlo approximation of $\overline{S}(\theta_n)$ is unbiased.
On Figure 1, for the three algorithms MCPG, SAPG and SAEM-pen, the evolution of an approximation of $\|S_{n+1} - \bar{S}(\theta_n)\|_2$ with iterations n is plotted, where, for a random variable U, $\|U\|_2 := \sqrt{\mathbb{E}\left[\|U\|^2\right]}$. This Monte Carlo approximation is computed from the 100 independent realizations of S_{n+1}; here, $\overline{S}(\theta_n)$ is explicit (see (20)). SAEM-pen and SAPG behave similarly; the L_2-norm converges to 0, and the convergence is slower when $\alpha = 0.6$ and $\beta = 0.1$ - this plot illustrates the result stated in Proposition 5, Section 4. This convergence does not hold for MCPG because the size m_n of the Monte Carlo approximation is kept fixed.
We compared the limiting vectors $\lim_n \theta_n$ obtained by each algorithm, over the 100 independent runs. They are all equal, and the limiting vector is also the limiting value θ_∞ of the EM-pen algorithm. In order to discuss the rate of convergence, we show the behavior of the algorithms when estimating the component #245

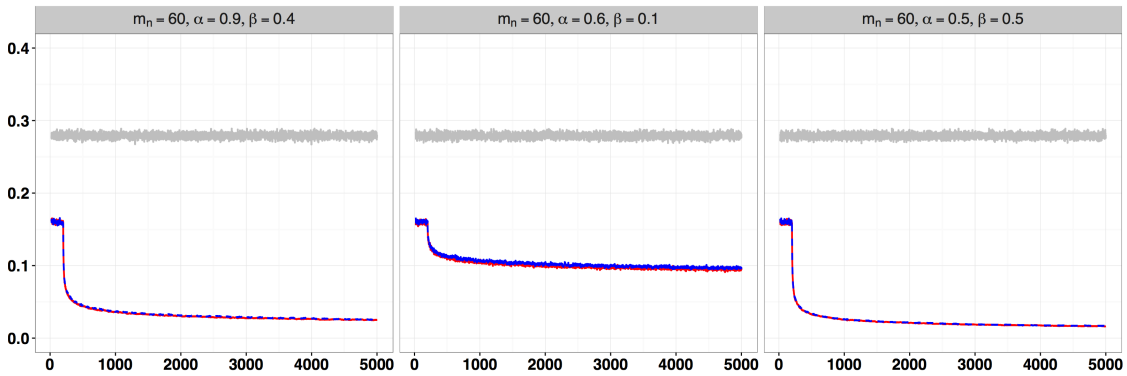

Figure 1: Evolution of the Monte Carlo approximation of $\|S_{n+1} - \bar{S}(\theta_n)\|_2$ with iterations n for algorithms MCPG (solid grey), SAEM-pen (dotdash red), SAPG (dashed blue), implemented with $(\alpha, \beta) = (0.9, 0.4)$ [left], $(\alpha, \beta) = (0.6, 0.1)$ [center] and $(\alpha, \beta) = (0.5, 0.5)$ [right]; for MCPG and SAPG, the batch size is fixed $m_n = 60$.

of the regression coefficients; this component was chosen among the non-null component of θ_∞. Figure 2 shows the boxplot of 100 estimations of the component #245 of the vector θ_n, when $n = 5, 25, 50, 500$ and 5000, for the algorithms MCPG, SAPG and SAEM-pen. Here, SAPG and MCPG behave similarly, with a smaller variability among the 100 runs than SAEM-pen. SAEM-pen converges faster than SAPG and MCPG which was expected since they correspond respectively to stochastic perturbations of EM-pen and GEM-pen algorithms. We observe also that the parameters α and β have a small influence on the rate of convergence.

Figure 3 shows the convergence of a Monte Carlo approximation of $\mathbb{E}[F(\theta_n)]$ based on 100 independent samples, along the iterations n for algorithms EM-pen, MCPG, SAPG and SAEM-pen; here, $(\alpha, \beta) = (0.9, 0.4)$ and $m_n = 60$. Here again, all the algorithms converge to the same value and EM-pen and SAEM-pen converge faster than MCPG and SAPG. We observe that the path of SAPG is far more smooth than the path of MCPG.

Finally, Figure 4 shows the support of the vector θ (where the component θ_1 and θ_{302} are removed) estimated by MCPG, SAPG, SAEM-pen and EM-pen. For each component, the number of times among 100 independent runs a component is in the support of the limit value $\lim_n \theta_n$, is displayed. Algorithms are implemented with $(\alpha, \beta) = (0.9, 0.4)$ and $m_n = 60$. For all algorithms, we observe that most of the non-null components of $\lim_n \theta_n$ are non-null components of $\theta^\star$. Note also that the stochastic algorithms MCPG, SAPG and SAEM-pen converge to the same vector as EM-pen.

3 Inference in non-linear mixed models for pharmacokinetic data

In this section, SAPG is applied to solve a more challenging problem. The application is in pharmacokinetic, with non-linear mixed effect models (NLMEM); in this application, the penalized maximum-likelihood inference is usually solved by the SAEM-pen algorithm, possibly combined with an approximation of the M-step when it is non explicit. This section also provides a numerical comparison of SAPG and SAEM-pen. Both algorithms have a simulation step; in this more challenging application, it will rely on a Markov chain Monte Carlo (MCMC) sampler - see Section 3.1. Therefore, for both algorithms, $\overline{S}(\theta)$ is approximated by a

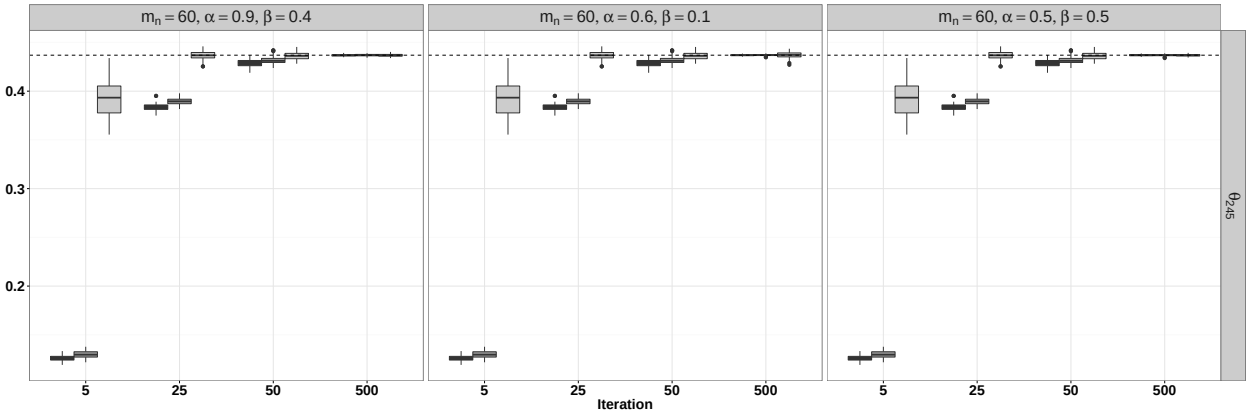

Figure 2: Evolution of the estimation of the component #245 of the vector θ along iterations n. MCPG (black), SAPG (dark grey) and SAEM-pen (light grey) are implemented with $(\alpha, \beta) = (0.9, 0.4)$ [left], $(\alpha, \beta) = (0.6, 0.1)$ [center] and $(\alpha, \beta) = (0.5, 0.5)$ [right]; for MCPG and SAPG, the batch size is fixed $m_n = 60$. Each boxplot is computed from 100 independent runs. Black dashed line correspond to the value obtained with EM-pen algorithm at iteration 5000.

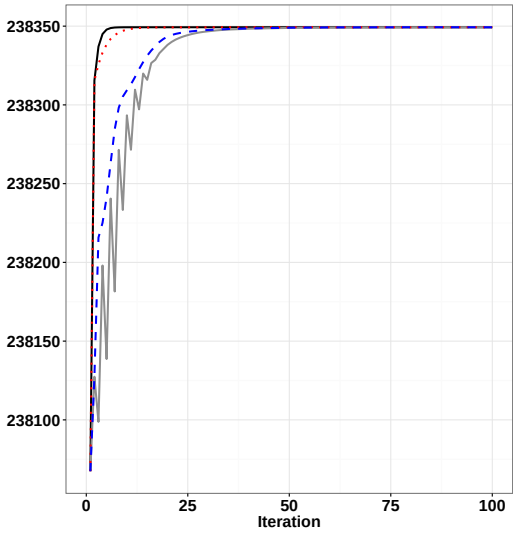

Figure 3: Monte Carlo approximation of $\mathbb{E}\left[F(\theta_n)\right]$ (based on 100 independent samples) along the iterations for algorithms EM-pen (solid black), MCPG (solid grey), SAEM-pen (dotdash red), SAPG (dashed blue), implemented with $(\alpha, \beta) = (0.9, 0.4)$ and $m_n = 60$.

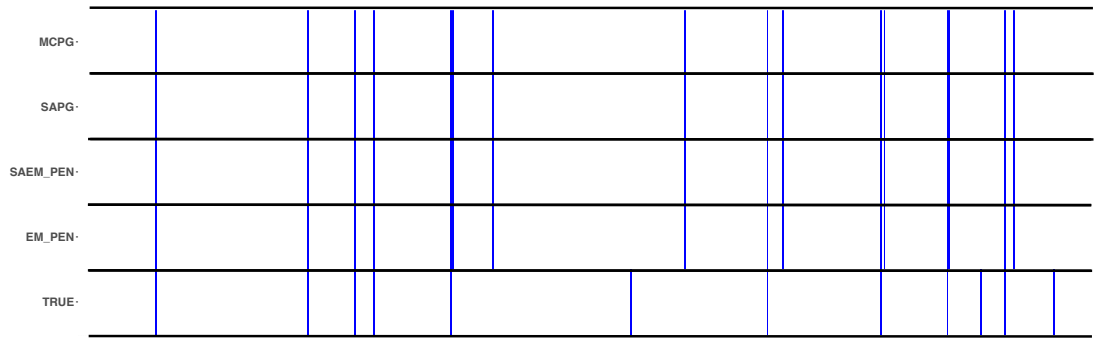

Figure 4: Support of $\lim_n \theta_n$ estimated - from top to bottom - with algorithms MCPG, SAPG, SAEM-pen and EM-pen over 100 runs for $(\alpha, \beta) = (0.9, 0.4)$ and $m_n = 60$. The support of $\theta_\star$ used to produce the observations is displayed on the bottom row. The components 1 and $D+1$ are not displayed.

biased Monte Carlo sum.
We start with a presentation of the statistical analysis and its translation into an optimization problem; we then propose a modification of the SAPG by allowing a random choice of the stepsize sequence $\{\gamma_n, n \geq 0\}$, to improve the numerical properties of the algorithm. We conclude the section by a comparison of the methods on a pharmacokinetic real data set.

3.1 The non-linear mixed effect model

Pharmacokinetic data are observed along time for N patients. Let Y_k be the vector of the J drug concentrations observed at time t_{kj} ($j \in \{1, \ldots, J\}$) for the k-th patient ($k \in \{1, \ldots, N\}$). The kinetic of the drug concentration is described by a non-linear pharmacokinetic regression model f, which is a function of time t and unobserved pharmacokinetic parameters $Z^{(k)}$. These parameters are typically the rates of absorption or elimination of the drug by the body. An example is detailed below. The variability among patients is modeled by the randomness of the hidden variables $Z^{(k)}$. These pharmacokinetic parameters may be influenced by covariates, such as age, gender but also genomic variables. Among these high dimension factors, only few of them are correlated to $Z^{(k)}$. Their selection can thus be performed by optimizing the likelihood with a sparsity inducing penalty, an optimization problem that enters problem (1). However, the likelihood is generally not concave, that is, through this example, we explore beyond the framework in which we are able to prove the convergence of MCPG and SAPG (see Section 4).
Let us now detail the model and the optimization problem. The mixed model is defined as

$$\mathsf{Y}_{kj} = f(t_{kj}, Z^{(k)}) + \epsilon_{kj}, \qquad \epsilon_{kj} \sim \mathcal{N}(0, \sigma^2) \text{ (iid)}, \tag{22}$$

where the measurement errors ϵ_{kj} are centered, independent and identically normally distributed with variance σ^2. Individual parameters $Z^{(k)}$ for the k-th subject is a R-dimensional random vector, independent of ϵ_{kj}. In a high dimension context, the $Z^{(k)}$'s depend on covariates (typically genomics variables) gathered in a matrix design $X_k \in \mathbb{R}^{R \times (D+1)R}$. The distribution of $Z^{(k)}$ is usually assumed to be normal with independent components

$$Z^{(k)} \overset{ind}{\sim} \mathcal{N}_R(X_k \mu, \Omega) \tag{23}$$

where $\mu \in \mathbb{R}^{(D+1)R}$ is the mean parameter vector and Ω is the covariance matrix of the random parameters $Z^{(k)}$, assumed to be diagonal. The unknown parameters are $\theta = \left(\mu, \Omega_{11}, \cdots, \Omega_{RR}, \sigma^2\right) \in \mathbb{R}^{R(D+1)} \times (0, +\infty)^{R+1}$.
A typical function f is the two-compartmental pharmacokinetic model with first order absorption, describing the distribution of a drug administered orally. The drug is absorbed from the gut and reaches the blood circulation where it can spread in peripheral tissues. This model corresponds to $f = \frac{A_c}{V_c}$ with A_c defined as

$$\begin{aligned} \frac{dA_d}{dt} &= -k_a\, A_d, \\ \frac{dA_c}{dt} &= k_a\, A_d + \frac{Q}{V_p} A_p - \frac{Q}{V_c} A_c - \frac{Cl}{V_c} A_c, \\ \frac{dA_p}{dt} &= \frac{Q}{V_c} A_c - \frac{Q}{V_p} A_p, \end{aligned} \tag{24}$$

with $A_d(0) = Dose$, $A_c(0) = 0$, $A_p(0) = 0$ and where A_d, A_c, A_p are the amount of drug in the depot, central and peripheral compartments, respectively; V_c, V_p are the volume of the central compartment and the peripheral compartment, respectively; Q and Cl are the inter compartment and global elimination clearances, respectively. To assure positiveness of the parameters, the hidden vector is

$$z = (\log(V_c), \log(V_p), \log(Q), \log(Cl), \log(k_a)).$$

It is easy to show that the model (22)-(23) belongs to the curved exponential family (4) with minimal sufficient statistics:

$$\begin{aligned} & S_{1k}(z) = z^{(k)}, \quad S_2(z) = \sum_{k=1}^{N} z^{(k)}\, z^{(k)\prime}, \\ & S_3(z) = \sum_{k=1}^{N} \sum_{j=1}^{J} (\mathsf{Y}_{kj} - f(t_{kj}, z^{(k)}))^2; \\ & \psi_{1k}(\theta) = (X_k \mu)' \Omega^{-1}, \quad \psi_2(\theta) = -\frac{1}{2} \Omega^{-1}, \\ & \psi_3(\theta) = -\frac{1}{2\sigma^2}, \end{aligned}$$

and $S(z) := \text{Vect}\,(S_{11}(z), \cdots, S_{1N}(z), S_2(z), S_3(z))$, $\psi := \text{Vect}\,(\psi_{11}, \cdots, \psi_{1N}, \psi_2, \psi_3)$. The function ϕ is given by $\phi(\theta) = -JN \log(\sigma) - \frac{N}{2} \log(|\Omega|) - \frac{1}{2} \sum_k (X_k \mu)' \Omega^{-1} (X_k \mu)$.
The selection of genomic variables that influence all coordinates of $Z^{(k)}$ could be obtained by optimizing the log-likelihood penalized by the function $g(\theta) = \lambda \|\mu\|_1$, the L_1 norm of μ with λ a regularization parameter. However, this estimator is not invariant under a scaling transformation (ie $\tilde{Z}^{(k)} = bZ^{(k)}$, $\tilde{\mu} = b\mu$ and $\tilde{\Omega}_{rr}^{1/2} = b\Omega_{rr}^{1/2}$) [Lehmann and Casella, 2006]. In our high dimension experiments, the scale of the hidden variables has a non negligible influence on the selection of the support. To be more precise, let us denote, for $r \in \{1, \ldots, R\}$,

$$\mu_{(r)} := (\mu_{(r-1)(D+1)+1}, \ldots, \mu_{r(D+1)})$$

the coordinates corresponding to the r-th pharmacokinetic parameter of function f. When the variance Ω_{rr} of the random parameters $Z_r^{(k)}$ is low, the algorithms tend to select too many covariates. This phenomenon

is strengthened with a small number of subjects as random effect variances are more difficult to estimate. A solution is to consider the following penalty

$$\lambda \sum_{r=1}^{R} \Omega_{rr}^{-\frac{1}{2}} \|\mu_{(r)}\|_1,$$

that makes the estimator invariant under scaling transformation. It was initially proposed by Städler et al. [2010] to estimate the regression coefficients and the residual error's variance in a mixture of penalized regression models. However, the resulting optimization problem is difficult to solve directly because the variance of the random effect Ω_{rr} appears in the penalty term. Therefore, we propose a new parametrisation

$$\tilde{\mu}_{(r)} := \mu_{(r)} \Omega_{rr}^{-\frac{1}{2}}, \qquad \Sigma_{rr} := \Omega_{rr}^{-\frac{1}{2}}$$

and $\tilde{\theta} := \{\tilde{\mu}, \Sigma_{11}, \cdots, \Sigma_{RR}, \sigma^2\} \in \mathbb{R}^{R(D+1)} \times (0, +\infty)^{R+1}$. Then, the optimization problem is the following:

$$\underset{\tilde{\theta}}{\text{Argmax}} \left(\ell(\tilde{\theta}) - g(\tilde{\theta}) \right), \quad \text{with } g(\tilde{\theta}) = \lambda \|\tilde{\mu}\|_1. \tag{25}$$

This problem can be solved using MCPG, SAPG or SAEM-pen algorithms. Indeed, the complete log-likelihood is now - up to an additive constant -

$$\begin{aligned} \log p(\mathsf{Y}, Z; \tilde{\theta}) = &- JN \log(\sigma) \\ &- \frac{1}{2} \sum_{k=1}^{N} \sum_{j=1}^{J} \frac{\left(Y_{kj} - f(t_{kj}, Z^{(k)})\right)^2}{\sigma^2} \\ &+ N \log(|\Sigma|) - \frac{1}{2} \sum_{k=1}^{N} \|\Sigma Z^{(k)} - X_k \tilde{\mu}\|^2 \end{aligned}$$

It is again a complete likelihood from the exponential family, with the statistic S unchanged and the functions ϕ and ψ given by - up to an additive constant -

$$\begin{aligned} &\phi(\tilde{\theta}) = -JN \log(\sigma) + N \log(|\Sigma|) - \frac{1}{2} \sum_{k=1}^{N} \|X_k \tilde{\mu}\|^2, \\ &\psi_{1k}(\tilde{\theta}) = \Sigma (X_k \tilde{\mu})^t, \quad \psi_2(\tilde{\theta}) = -\frac{1}{2} \Sigma^2, \quad \psi_3(\tilde{\theta}) = -\frac{1}{2\sigma^2}. \end{aligned}$$

With these definitions of ϕ, ψ and g, the M-step of SAEM-pen amounts to compute the optimum of a convex function, which is solved numerically by a call to a cyclical coordinate descent implemented in the R package glmnet [Friedman et al., 2010].

MCMC sampler. In the context of non-linear mixed models, simulation from $\pi_{\theta_n} \mathrm{d}\mu$ can not be performed directly like in the toy example. We then use a MCMC sampler based on a Metropolis Hastings algorithm to perform the simulation step. Two proposal kernels are successively used during the iterations of the Metropolis Hastings algorithm. The first kernel corresponds to the prior distribution of $\Sigma Z^{(k)}$ that is the Gaussian distribution $\mathcal{N}(X_k \tilde{\mu}_n, I)$. The second kernel corresponds to a succession of R uni-dimensional random walk in order to update successively each component of $Z^{(k)}$. The variance of each random walk is automatically tuned to reach a target acceptance ratio following the principle of an adaptive MCMC algorithm [Andrieu and Thoms, 2008].

Adaptive random stepsize sequences. Numerical experiments reveal that choosing a deterministic sequence $\{\gamma_n, n \geq 0\}$ yields to instability of the SAPG algorithm. Indeed, in the context of NLMEM, parameters to estimate are of different scales. For example, random effect and residual variances are constrained to be positive. Some of them are close to zero, some are not. Then large values of the gradient make the algorithm diverges if the sequence $\{\gamma_n, n \geq 0\}$ decreases too slowly. It can be time consuming and hand-tuning to find a sequence that stabilizes the algorithm.
As an alternative, we propose to use a matrix-valued random sequence $\{\Gamma_n, n \geq 0\}$ and replace the update rule of SAPG by

$$\theta_{n+1} = \mathrm{Prox}_{\gamma_{n+1} g}\left(\theta_n + \Gamma_{n+1}\left(\phi(\theta_n) + \Psi(\theta_n) S_{n+1}^{\mathrm{sa}}\right)\right).$$

The matrix Γ_{n+1} is defined as follows. First note that the gradient and the hessian of the likelihood $\ell(\theta)$ can be approximated by stochastic approximation using the Louis principle [see McLachlan and Krishnan, 2008, Chapter 4]. Let us denote H_n the stochastic approximation of the hessian obtained at iteration n as explained by Samson et al. [2007]. Note that no supplementary random samples are required to obtain this approximation. Then Γ_{n+1} is defined as a diagonal matrix with entries $(\Gamma_{n+1})_{ii} = 1/(H_n)_{ii}$. Along the iterations, each diagonal entry of the matrix Γ_{n+1} converges: this limiting value can be seen as a simple way to automatically tune a good $\gamma_\star$, that is parameter specific. Then (say, after iteration n_0) the entries $(\Gamma_{n+1})_{ii}$ are decreased progressively to make the algorithm converges. Hence, Γ_{n+1} is chosen as follows

$$(\Gamma_{n+1})_{ii} = \begin{cases} 1/(H_n)_{ii} & \text{if } n \leq n_0, \\ ((n-n_0)^\alpha (H_n)_{ii})^{-1} & \text{if } n > n_0. \end{cases} \tag{26}$$

3.2 Simulated data set.

The convergence of the corresponding algorithms is illustrated on simulated data. Data are generated with model (24) and $N = 40$, $J = 12$, $D = 300$. The design matrix X_k is defined by Eq. (18), with components $(X_{k1}, \ldots, X_{kD})$ drawn from $\mathcal{N}(0, \Gamma)$ with $\Gamma_{ii'} = 0.5^{|i-i'|}$ $(i, i' = 1, ..., 300)$. Parameter values are

$$\begin{aligned} &[\mu_1, \mu_{1+(D+1)}, \mu_{1+2(D+1)}, \mu_{1+3(D+1)}, \mu_{1+4(D+1)}] \\ &= [5.5, 5.84, 7.03, 6.68, -0.43]; \end{aligned}$$

the other components are set to zero, except μ_4 and μ_{912} that are set to 1. The matrix Ω is diagonal with diagonal elements equal to $(0.16, 0.16, 0.16, 0.04, 0.04)$.
The penalty function is set to

$$g(\tilde{\theta}) := \lambda \sum_{\ell \neq \{1+r(D+1), r=0,\cdots,4\}} |\tilde{\mu}_\ell|, \tag{27}$$

only the parameters corresponding to a covariate effect being penalized. The optimization problem (1) with regularization parameter $\lambda = 190$ is solved on this dataset with SAEM-pen and SAPG; we run SAPG with the random sequence $\{\Gamma_n, n \geq 0\}$ as described above (see (26)) with $n_0 = 3500$. For both algorithms, the stochastic approximation step size was set to:

$$\delta_{n+1} = \begin{cases} 0.75 & \text{if } n \leq n_0 \\ \frac{0.75}{(n-n_0)^\alpha} & \text{if } n > n_0 \end{cases} \tag{28}$$

We set $\alpha = 0.75$ and $\beta = 0.499$. Figure 5 shows the convergence of SAEM-pen and the modified SAPG for the estimation of each component of θ. The two algorithms have similar convergence behavior. All the parameters corresponding to a covariate effect are estimated to zero except the two components μ_4 and μ_{912}. Figure 6 presents the evolution of four entries of the matrix Γ_n along the iterations of SAPG, corresponding to the components $\tilde{\mu}_{904}$, $\tilde{\mu}_{912}$, Σ_{44} and σ. We can notice that they are not on the same scale. They vary

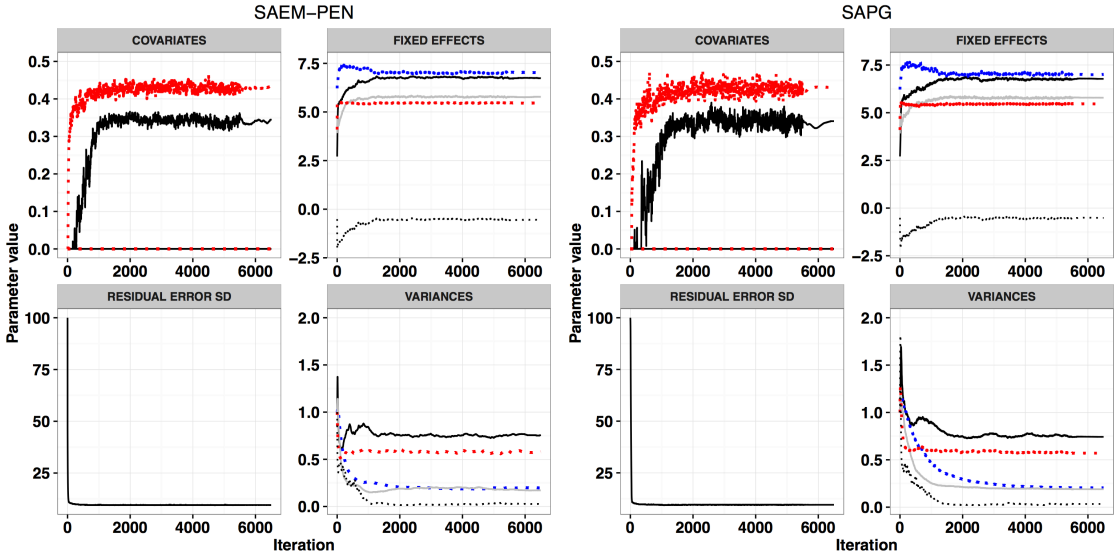

Figure 5: A path of SAEM-pen [left part of the plot] and SAPG run with a random sequence $\{\Gamma_n, n \geq 0\}$ [right part of the plot]. For each algorithm, estimation of the standard deviation of the residual error σ [bottom left]; the variances of the $Z^{(k)}$'s, $\Omega_{11}, \cdots, \Omega_{RR}$ [bottom right]; the path of the covariate parameters μ_i for $i \notin \{1, 1+(D+1), \cdots, 1+4(D+1)\}$ [top left]; the path of the intercept parameters $\mu_i, i \in \{1, 1+(D+1), \cdots, 1+4(D+1)\}$ [top right]. Each color and line type corresponds to a specific parameter: red dotted line for Cl, blue dotted for V_c, black dotted line for k_a, grey solid line for Q and yellow for V_p. Note that the path of all the covariate parameters is zero except for two components.

during the first iterations and converge to limiting values before iteration $n_0 = 3500$. Then the step sizes decrease to 0, following the definition given in (26).

3.3 Application to real data

Algorithms SAEM-pen and SAPG with matrix-valued random sequence $\{\Gamma_n, n \geq 0\}$ are applied to real data of the pharmacokinetic of dabigatran (DE) from two cross over clinical trials [Delavenne et al., 2013, Ollier et al., 2015]. These 2 trials studied the drug-drug interaction between DE and different Pgp-inhibitors. From these 2 trials, the pharmacokinetics of DE are extracted from 15 subjects with no concomitant treatment with Pgp-inhibitors. The concentration of dabigatran is measured at 9 sampling times for each patient. Each subject is genotyped using the DMET® microarray from Affymetrix. Single nucleotide polymorphisms (SNP) showing no variability between subjects are removed and 264 SNP are included in the analysis.
Function f of the non-linear mixed model is defined as the two compartment pharmacokinetic model with first order absorption previously described (see (24)) [Delavenne et al., 2013]. The penalty function g is defined by (27).
Because of the limited number of subjects, the influence of genetic covariates is only studied on V_c and Cl parameters, that characterize the elimination process and are the most likely to be influenced by the genetic. Finally, random effect variances of Q and V_p are set to 0.01 in accordance with previously published population pharmacokinetic of dabigatran [Delavenne et al., 2013]. The other variance parameters are

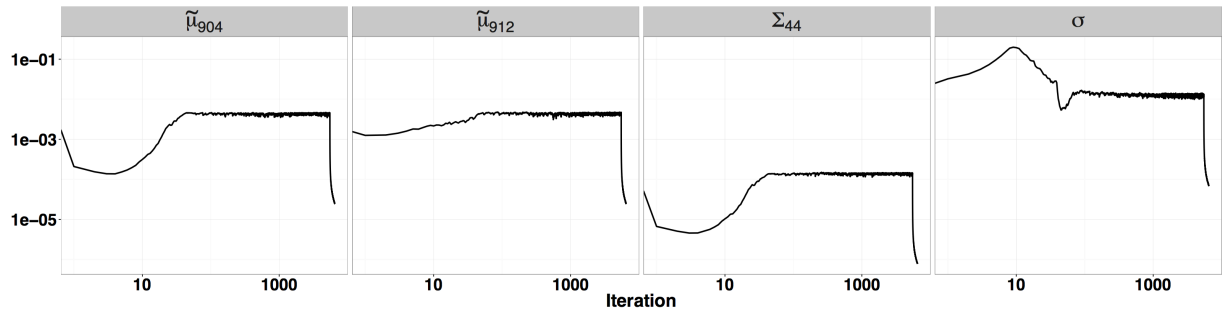

Figure 6: Evolution of $(\Gamma_n)_{ii}$ with iterations n of SAPG, for four different values of i, corresponding to the components $\tilde{\mu}_{904}$ [left]; $\tilde{\mu}_{912}$ [middle left]; Σ_{44} [middle right]; σ [right]. Both x-axis and y-axis are in $\log_{10}$ scale.

estimated. The penalized likelihood problem (25) is solved on the data with the SAEM-pen and SAPG algorithms, for 40 different values of parameter λ. SAPG algorithm is run using the random sequence $\{\Gamma_n, n \geq 0\}$ given in Eq. (26). The best regularization parameter λ is chosen with a data-driven approach based on the EBIC criteria [Chen and Chen, 2008].
Figure 7 shows the results. The regularization paths of Cl and V_c parameters using both algorithms correspond to the evolution of covariate coefficient estimates as a function of the value of λ. They are reconstructed with low noise for both algorithms, are very similar for high values of λ but less for lower values of λ.
Finally, the selected model has all covariates parameters set to zero. This means that none of the genetic covariates influence the distribution of the individual parameters. This result is not surprising given the low number of subjects and the fact that a large part of the inter individual variability is due to the dissolution process of the drug [Ollier et al., 2015] and is therefore not influenced by genetic covariates. This lack of relationship between dabigtran's pharmacokinetic parameters and genetic covariates has already been highlighted in an other study [Gouin-Thibault et al., 2016].

4 Convergence of MCPG and SAPG

The convergence of MCPG and SAPG is established by applying recent results on the convergence of perturbed Proximal-Gradient algorithms (Atchadé et al. [2016]). A slight adaptation of [Atchadé et al., 2016, Theorem 2] to the case $\nabla\ell(\theta)$ is of the form $\nabla\phi(\theta) + \Psi(\theta)\overline{S}(\theta)$, where $\overline{S}(\theta)$ is an intractable expectation and $\nabla\phi, \Psi$ are explicit, yields

Theorem 2. *Assume H1, H2, $\theta \mapsto \ell(\theta)$ is concave, and the set $\mathcal{L} := \operatorname{argmax}_{\theta\in\Theta} F(\theta)$ is a non empty subset of Θ. Let $\{\theta_n, n \geq 0\}$ be given by*

$$\theta_{n+1} = \operatorname{Prox}_{\gamma_{n+1},g}\left(\theta_n + \gamma_{n+1}\left\{\nabla\phi(\theta_n) + \Psi(\theta_n)S_{n+1}\right\}\right),$$

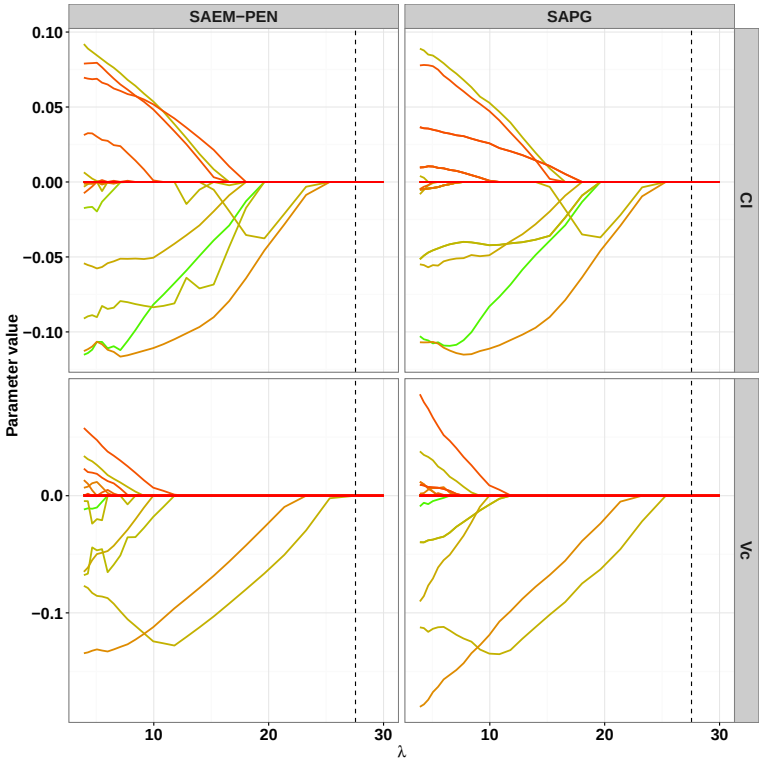

Figure 7: Regularization path of covariate parameters (Cl parameter on top, V_c parameter on bottom) obtained on dabigatran pharmacokinetic data for both SAEM-pen and SAPG algorithms. Black vertical dashed line corresponds to the λ value selected by $EBIC$.

with a $(0, 1/L]$*-valued stepsize sequence* $\{\gamma_n, n \geq 0\}$ *satisfying* $\sum_n \gamma_n = +\infty$*. If the series*

$$\sum_n \gamma_{n+1} \left\langle \Psi(\theta_n) \left(S_{n+1} - \overline{S}(\theta_n)\right), T_{\gamma_{n+1}}(\theta_n) \right\rangle,$$
$$\sum_n \gamma_{n+1} \Psi(\theta_n) \left(S_{n+1} - \overline{S}(\theta_n)\right),$$
$$\sum_n \gamma_{n+1}^2 \|\Psi(\theta_n) \left(S_{n+1} - \overline{S}(\theta_n)\right)\|^2,$$

converge, where

$$T_\gamma(\theta) := \mathrm{Prox}_{\gamma,g}(\theta + \gamma \left\{\nabla\phi(\theta) + \Psi(\theta)\overline{S}(\theta)\right\}),$$

then there exists $\theta_\infty \in \mathcal{L}$ *such that* $\lim_n \theta_n = \theta_\infty$*.*

We check the conditions of Theorem 2 in the case S_{n+1} is resp. given by (11) for the proof of MCPG and by (12) for the proof of SAPG. Our convergence analysis is restricted to the case ℓ is concave; to our best knowledge, the convergence of the perturbed Proximal-Gradient algorithms when ℓ is not concave is an open question.
The novelty in this section is Proposition 5 and Theorem 6 which provide resp. a control of the L_2-norm of the error $S_{n+1}^{\mathrm{sa}} - \overline{S}(\theta_n)$ and the convergence of SAPG. These results necessitate a control of $\left(S_{n+1}^{\mathrm{sa}} - \overline{S}(\theta_n)\right)$ when the approximation S_{n+1}^{sa} (see Eq. (12)) is a weighted sum of the function S evaluated at all the samples $\{Z_{i,j}, i \leq m_{j+1}, j \leq n\}$ drawn from the initialization of the algorithm. This approximation differs from a more classical Monte Carlo approximation (see Theorems 3 and 4 for the convergence of MCPG, results directly adapted from Atchadé et al. [2016]).

We allow the simulation step of MCPG and SAPG to rely on Markov chain Monte Carlo sampling: at iteration $(n+1)$, the conditional distribution of $Z_{j+1,n}$ given the past is $P_{\theta_n}(Z_{j,n}, \cdot)$ where P_θ is a Markov transition kernel having $\pi_\theta \mathrm{d}\mu$ as unique invariant distribution. The control of the quantities $S_{n+1} - \overline{S}(\theta_n)$ requires some ergodic properties on the kernels $\{P_{\theta_n}, n \geq 0\}$ along the path $\{\theta_n, n \geq 0\}$ produced by the algorithm. These properties have to be uniform in θ, a property often called, after Roberts and Rosenthal [2007], the "containment condition" (see e.g. the literature on the convergence of adaptive MCMC samplers, for example Andrieu and Moulines [2006], Roberts and Rosenthal [2007], Fort et al. [2011b]). There are therefore three main strategies to prove the containment condition. In the first strategy, Θ is assumed to be bounded, and a uniform ergodic assumption on the kernels $\{P_\theta, \theta \in \Theta\}$ is assumed. In the second one, there is no boundedness assumption on Θ but the property $\mathbb{P}(\limsup_n \|\theta_n\| < \infty) = 1$ has to be established prior the proof of convergence; a kind of local boundedness condition on the sequence $\{\theta_n, n \geq 0\}$ is then applied - see e.g. Andrieu and Moulines [2006], Fort et al. [2011b]. The last strategy consists in showing that $\mathbb{P}(\sup_n \rho_n \|\theta_n\| < \infty) = 1$ for some deterministic sequence $\{\rho_n, n \geq 0\}$ vanishing to zero when $n \to \infty$ at a rate compatible with the decaying ergodicity rate - see e.g. Saksman and Vihola [2010]. The last two strategies are really technical and require from the reader a strong background on controlled Markov chain theory; for pedagogical purposes, we therefore decided to state our results in the first context: we will assume that Θ is bounded.
By allowing MCMC approximations, we propose a theory which covers the case of biased approximation, called below the *biased case*: conditionally to the past

$$\mathcal{F}_n := \sigma\left(Z_{i,j}, i \leq m_{j+1}, j \leq n-1\right), \tag{29}$$

the expectation of S_{n+1} is not $\overline{S}(\theta_n)$: $\mathbb{E}\left[S_{n+1}|\mathcal{F}_n\right] \neq \overline{S}(\theta_n)$. As soon as the samplers $\{P_\theta, \theta \in \Theta\}$ are ergodic enough (for example, under H4a) and H4b)), the bias vanishes when the number of Monte Carlo points

m_n tends to infinity. Therefore, the proof for the biased case when the sequence $\{m_n, n \geq 0\}$ is constant is the most technical situation since the bias does not decay. It relies on a specific decomposition of the error $S_{n+1} - \overline{S}(\theta_n)$ into a martingale increment with bounded L^2-moments, and a remainder term which vanishes when $n \to \infty$ even when the batch size m_n is constant. Such a behavior of the remainder term is a consequence of regularity properties on the functions $\nabla\phi$, Ψ, $\overline{S}$ (see H3c)), on the proximity operator (see H3d)) and on the kernels $\{P_\theta, \theta \in \Theta\}$ (see H4c)).
Our theory also covers the *unbiased case* i.e. when

$$\mathbb{E}\left[S_{n+1}|\mathcal{F}_n\right] = \overline{S}(\theta_n)$$

.

We therefore establish the convergence of MCPG and SAPG by strengthening the conditions H1 and H2 with

H3. *a) ℓ is concave and the set $\mathcal{L} := \operatorname{argmax}_\Theta F$ is a non-empty subset of Θ.*

b) Θ is bounded.

c) There exists a constant L such that for any $\theta, \theta' \in \Theta$,

$$\|\nabla\phi(\theta) - \nabla\phi(\theta')\| + \|\Psi(\theta) - \Psi(\theta')\| + \|\overline{S}(\theta) - \overline{S}(\theta')\| \leq L\|\theta - \theta'\|,$$

where for a matrix A, $\|A\|$ denotes the operator norm associated with the Euclidean vector norm.

d) $\sup_{\gamma\in(0,1/L]} \sup_{\theta\in\Theta} \gamma^{-1}\|\operatorname{Prox}_{\gamma,g}(\theta) - \theta\| < \infty$.

Note that the assumptions H3b)-H3c) imply (3) and $\sup_{\theta\in\Theta}\left(\|\nabla\phi(\theta)\| + \|\Psi(\theta)\| + \|\overline{S}(\theta)\|\right) < \infty$. When Θ is a compact convex set, then H3d) holds for the elastic net penalty, the Lasso or the fused Lasso penalty. [Atchadé et al., 2016, Proposition 11] gives general conditions for H3d) to hold.

Before stating the ergodicity conditions on the kernels $\{P_\theta, \theta \in \Theta\}$, let us recall some basic properties on Markov kernels. A Markov kernel P on the measurable set $(\mathsf{Z}, \mathcal{Z})$ is an application on $\mathsf{Z} \times \mathcal{Z}$, taking values in $[0, 1]$ such that for any $x \in \mathsf{Z}$, $P(x, \cdot)$ is a probability measure on $\mathcal{Z}$; and for any $A \in \mathcal{Z}$, $x \mapsto P(x, A)$ is measurable. Furthermore, if P is a Markov kernel, P^k denotes the k-th iterate of P defined by induction as

$$\begin{aligned} P^0(x, A) &:= \mathbb{1}_A(x), \\ P^k(x, A) &:= \int P^{k-1}(x, \mathrm{d}z)P(z, A), \qquad k \geq 1. \end{aligned}$$

Finally, the kernel P acts on the probability measures: for any probability measure ξ on $\mathcal{Z}$, ξP is a probability measure defined by

$$\xi P(A) := \int \xi(\mathrm{d}z)P(z, A), \qquad A \in \mathcal{Z};$$

and P acts on the positive measurable functions: for a measurable function $f : \mathsf{Z} \to \mathbb{R}_+$, Pf is a measurable function defined by

$$Pf(z) := \int f(y)\, P(z, \mathrm{d}y).$$

We refer the reader to Meyn and Tweedie [2009] for the definitions and basic properties on Markov chains. Given a measurable function $W : \mathsf{Z} \to [1, +\infty)$, define the W-norm of a signed measure μ on $\mathcal{Z}$ and the W-norm of a function $f : \mathsf{Z} \to \mathbb{R}^d$:

$$|f|_W := \sup_{\mathsf{Z}} \frac{\|f\|}{W}, \qquad \|\mu\|_W := \sup_{f:|f|_W \leq 1} \left| \int f \mathrm{d}\mu \right|;$$

these norms generalize resp. the supremum norm of a function and the total variation norm of a measure. Our results are derived under the following conditions on the kernels:

H4. *a) There exist $\lambda \in (0, 1]$, $b < \infty$ and a measurable function $W : \mathsf{Z} \to [1, +\infty)$ such that*

$$|S|_{\sqrt{W}} < \infty, \qquad \sup_{\theta \in \Theta} P_\theta W \leq \lambda W + b.$$

b) There exist constants $C < \infty$ and $\rho \in (0, 1)$ such that for any $z \in \mathsf{Z}$ and $n \geq 0$,

$$\sup_{\theta \in \Theta} \|P_\theta^n(z, \cdot) - \pi_\theta\|_W \leq C \, \rho^n \, W(z).$$

c) There exists a constant C such that for any $\theta, \theta' \in \Theta$,

$$\|\pi_\theta - \pi_{\theta'}\|_{\sqrt{W}} + \sup_{z \in \mathsf{Z}} \frac{\|P_\theta(z, \cdot) - P_{\theta'}(z, \cdot)\|_{\sqrt{W}}}{\sqrt{W}(z)} \leq C \, \|\theta - \theta'\|.$$

Sufficient conditions for the uniform-in-θ ergodic behavior H4b) are given e.g. in [Fort et al., 2011a, Lemma 2.3.]. Examples of MCMC kernels P_θ satisfying these assumptions can be found in [Andrieu and Moulines, 2006, Proposition 12], [Saksman and Vihola, 2010, Proposition 15], [Fort et al., 2011a, Proposition 3.1.], [Schreck et al., 2013, Proposition 3.2.] and [Fort et al., 2015, Proposition 3.1.].

Theorem 3 establishes the convergence of MCPG when the number of points in the Monte Carlo sum S_{n+1}^{mc} is constant over iterations and the step size sequence $\{\gamma_n, n \geq 0\}$ vanishes at a convenient rate. It is proved in [Atchadé et al., 2016, Theorem 4].

Theorem 3. *Assume H1, H2, H3a-c) and H4a-b). Let $\{\theta_n, n \geq 0\}$ be the sequence given by Algorithm 1 with a $(0, 1/L]$-valued sequence $\{\gamma_n, n \geq 0\}$ such that $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 < \infty$, and with a constant sequence $\{m_n, n \geq 0\}$.*
In the biased case, assume also H3d) and H4c) and $\sum_n |\gamma_{n+1} - \gamma_n| < \infty$.
Then, with probability one, there exists $\theta_\infty \in \mathcal{L}$ such that $\lim_n \theta_n = \theta_\infty$.

Theorem 4 establishes the convergence of MCPG when the number of points in the Monte Carlo sum S_{n+1}^{mc} is increasing; it allows a constant stepsize sequence $\{\gamma_n, n \geq 0\}$. It is proved in [Atchadé et al., 2016, Theorem 6].

Theorem 4. *Assume H1, H2, H3a-c) and H4a-b). Let $\{\theta_n, n \geq 0\}$ be the sequence given by Algorithm 1 with a $(0, 1/L]$-valued sequence $\{\gamma_n, n \geq 0\}$ and an integer valued sequence $\{m_n, n \geq 0\}$ such that $\sum_n \gamma_n = +\infty$ and $\sum_n \gamma_n^2 / m_n < \infty$.*
In the biased case, assume also $\sum_n \gamma_n / m_n < \infty$.
Then, with probability one, there exists $\theta_\infty \in \mathcal{L}$ such that $\lim_n \theta_n = \theta_\infty$.

MCPG and SAPG differ from the approximation of $\overline{S}(\theta_n)$ at each iteration. We provide below a control of this error for a constant or polynomially increasing batch size sequence $\{m_n, n \geq 0\}$ and polynomially decreasing stepsize sequences $\{\gamma_n, n \geq 0\}$ and $\{\delta_n, n \geq 0\}$.

Proposition 5. *Let $\gamma_\star, \delta_\star, m_\star$ be positive constants and $\beta \in [0,1)$, $\alpha \geq \beta$, $c \geq 0$. Set $\gamma_n = \gamma_\star n^{-\alpha}$, $\delta_n = \delta_\star n^{-\beta}$ and $m_n = m_\star n^c$. Assume H1 to H4. Then*

$$\mathbb{E}\left[\|S_{n+1}^{\mathrm{mc}} - \overline{S}(\theta_n)\|^2\right] = O\left(n^{-c}\right),$$
$$\mathbb{E}\left[\|S_{n+1}^{\mathrm{sa}} - \overline{S}(\theta_n)\|^2\right] = O\left(n^{-\{2(\alpha-\beta)\wedge(\beta+c)\}}\right).$$

The proof is given in Section 5.3. This proposition shows that when applying MCPG with a constant batch size ($c = 0$), the error $S_{n+1}^{\mathrm{mc}} - \overline{S}(\theta_n)$ does not vanish; this is not the case for SAPG, since even when $c = 0$, the error $S_{n+1}^{\mathrm{sa}} - \overline{S}(\theta_n)$ vanishes as soon as $\alpha > \beta > 0$. Since the case "constant batch size" is the usual choice of the practitioners in order to reduce the computational cost of the algorithm, this proposition supports the use of SAPG instead of MCPG.
We finally study the convergence of SAPG without assuming the batch size sequence $\{m_n, n \geq 0\}$ is constant, which implies the following assumption on the sequences $\{\gamma_n, \delta_n, m_n, n \geq 0\}$.

H5. *The step size sequences $\{\gamma_n, n \geq 0\}$, $\{\delta_n, n \geq 0\}$ and the batch size sequence $\{m_n, n \geq 0\}$ satisfy*

a) $\gamma_n \in (0, 1/L]$, $\delta_n \in (0,1)$, $m_n \in \mathbb{N}$, $\sum_n \gamma_n = +\infty$, $\sum_n \gamma_n^2 < \infty$,

$$\sum_n \left(\gamma_{n-1}\gamma_n + \gamma_{n-1}^2 + |\gamma_n - \gamma_{n-1}|\right) \mathsf{D}_n < \infty,$$
$$\sum_n \gamma_n^2 \delta_n^2 (1 + \mathsf{D}_{n+1})^2 m_n^{-1} < \infty,$$

where $\mathsf{D}_n := \sum_{k \geq n} \left(\prod_{j=n}^k (1 - \delta_j)\right)$.

b) Furthermore,

$$\sum_n \gamma_{n+1} |m_{n+1}^{-1}\delta_{n+1} - m_n^{-1}\delta_n| < \infty,$$
$$\sum_n \gamma_{n+1} |m_{n+1}^{-1}\delta_{n+1}\mathsf{D}_{n+2} - m_n^{-1}\delta_n \mathsf{D}_{n+1}| < \infty,$$
$$\sum_n \left(\gamma_{n-1}\gamma_n + \gamma_{n-1}^2 + |\gamma_n - \gamma_{n-1}|\right) \cdots$$
$$\times\, m_{n-1}^{-1}\, \delta_{n-1}(1 + \mathsf{D}_n) < \infty.$$

We comment this assumption in the case the batch size sequence $\{m_n, n \geq 0\}$ is constant, since it is the usual choice of the practitioners. When $\delta_n = \delta_\star \in (0,1)$ for any $n \geq 0$, then $\mathsf{D}_n = \delta_\star^{-1}$ for any $n \geq 0$. This implies that the condition H5 is satisfied with polynomially decreasing sequences $\gamma_n \sim \gamma_\star / n^\alpha$ with $\alpha \in (1/2, 1]$ (and $m_n = m$ for any n).
When $\delta_n \sim \delta_\star\, n^{-\beta}$ for $\beta \in (0,1)$, then $\mathsf{D}_n = O(n^\beta)$ (see Lemma 9). Hence, using Lemma 9, H5a) and H5b) are satisfied with $\gamma_n \sim \gamma_\star n^{-\alpha}$ where $\beta < (1+\beta)/2 < \alpha \leq 1$, and $m_n = m$ for any n.
We can not have $\delta_n = \delta_\star n^{-1}$ since it implies $\mathsf{D}_n = +\infty$ for any $n \geq 0$.

Theorem 6. *Assume H1, H2, H3 and H4a-b).*
Let $\{\theta_n, n \geq 0\}$ be the sequence given by Algorithm 2 and applied with sequences $\{\gamma_n, \delta_n, m_n, n \geq 0\}$ verifying H5a).
In the biased case, assume also H4c) and H5b).
Then with probability one, there exists $\theta_\infty \in \mathcal{L}$ such that $\lim_n \theta_n = \theta_\infty$.

Proof. The proof is in Section 5.4 □

5 Proof

5.1 Proof of Proposition 1

Lemma 7. *Under the assumptions of Proposition 1, for any $\gamma \in (0, 1/L]$, $s \in \overline{S}(\Theta)$ and any $\theta, \theta' \in \Theta$,*

$$\mathcal{Q}(\theta|\theta') \geq \mathcal{Q}(\theta'|\theta') - \frac{1}{2\gamma}\|\theta - \theta' - \gamma\{\nabla\phi(\theta') + \mathrm{J}\,\psi(\theta')\overline{S}(\theta')\}\|^2 + \frac{\gamma}{2}\|\nabla\phi(\theta') + \mathrm{J}\,\psi(\theta')\overline{S}(\theta')\|^2.$$

Proof. Fix $\theta' \in \Theta$ and $s \in \overline{S}(\Theta)$. The derivative of the function $\theta \mapsto \mathcal{L}(\theta) := \phi(\theta) + \langle s, \psi(\theta)\rangle$ is $\nabla\phi(\theta) + \mathrm{J}\,\psi(\theta)s$ and this gradient is L-lipschitz. From a Taylor expansion to order 1 at θ' and since the gradient is Lipschitz, we have

$$\mathcal{L}(\theta) \geq \mathcal{L}(\theta') + \langle \nabla\phi(\theta') + \mathrm{J}\,\psi(\theta')s, \theta - \theta'\rangle - \frac{L}{2}\|\theta - \theta'\|^2.$$

We then choose $s = \overline{S}(\theta')$, use $L \leq 1/\gamma$ and conclude by the equality $2\langle a, b\rangle - \|a\|^2 = \|b\|^2 - \|a - b\|^2$. □

Proof of Proposition 1 We prove that $\mathcal{Q}(\theta_{n+1}|\theta_n) - g(\theta_{n+1}) \geq \mathcal{Q}(\theta_n|\theta_n) - g(\theta_n)$ so that the sequence $\{\theta_n, n \geq 0\}$ defined by (10) is a sequence satisfying (15).
By Lemma 7, it holds for any $\theta \in \Theta$ and any $\gamma \in (0, 1/L]$

$$\begin{aligned}\mathcal{Q}(\theta|\theta_n) - g(\theta) \\ &\geq \mathcal{Q}(\theta_n|\theta_n) + \frac{\gamma}{2}\|\nabla\phi(\theta_n) + \mathrm{J}\,\psi(\theta_n)\overline{S}(\theta_n)\|^2 \\ &\qquad - \frac{1}{2\gamma}\left\|\theta - \theta_n - \gamma\{\nabla\phi(\theta_n) + \mathrm{J}\,\psi(\theta_n)\overline{S}(\theta_n)\}\right\|^2 - g(\theta).\end{aligned}$$

Note that the RHS and the LHS are equal when $\theta = \theta_n$ so that for any point τ which maximizes the RHS, it holds $\mathcal{Q}(\tau|\theta_n) - g(\tau) \geq \mathcal{Q}(\theta_n|\theta_n) - g(\theta_n)$. This concludes the proof upon noting that such a point τ is unique and equal to θ_{n+1} given by (10).

5.2 Technical lemmas

Define

$$\Delta_{k:n} := \prod_{j=k}^{n}(1 - \delta_j), \quad 0 \leq k \leq n, \quad \Delta_{n+1:n} = 1,$$
$$\mathsf{D}_k := \sum_{n \geq k} \Delta_{k:n}.$$

Lemma 8. *For any $n \geq 2$, $\sum_{j=2}^{n} \Delta_{j+1:n}\,\delta_j = 1 - \Delta_{2:n}$.*

Proof. For any $j \leq n$, we have $\Delta_{j+1:n} - \Delta_{j:n} = \delta_j \Delta_{j+1:n}$ from which the result follows. $\square$

Lemma 9. *Let $\beta \in (0,1)$ and $\delta_\star > 0$. Set $\delta_n = \delta_\star n^{-\beta}$ for any $n \geq 1$. Then for any k large enough,*

$$\delta_k \mathsf{D}_k \leq 1 + O\left(k^{\beta-1}\right).$$

Furthermore, $|\delta_{n+1}\mathsf{D}_{n+2} - \delta_n \mathsf{D}_{n+1}| = O(1/n^{1+(1-\beta)\wedge\beta})$.

The proof of Lemma 9 relies on standard Taylor's expansions with explicit formulas for the remainder. The proof is omitted.

Lemma 10. *Let $\beta \in [0,1)$ and $\delta_\star > 0$. For any r, when $n \to \infty$,*

$$\sum_{j=2}^{n} j^{-r} \prod_{k=j}^{n} \left(1 - \frac{\delta_\star}{k^\beta}\right) = O\left(n^{\beta-r}\right).$$

Proof. We have

$$\begin{aligned}
\prod_{k=j}^{n} \left(1 - \frac{\delta_\star}{k^\beta}\right) &\leq \exp\left(-\delta_\star \sum_{k=j}^{n} k^{-\beta}\right) \\
&\leq \exp\left(-\frac{\delta_\star}{1-\beta}\left\{n^{1-\beta} - j^{1-\beta}\right\}\right).
\end{aligned}$$

Let $q_\star \geq 0$ such that for any $q \geq q_\star$, $q(1-\beta) + 1 - r > 0$. For any constant $D > 0$, there exist constants C, C' (whose value can change upon each appearance) such that

$$\begin{aligned}
&\sum_{j=2}^{n} j^{-r} \exp(Dj^{1-\beta}) \\
&= \sum_{q \geq 0} \frac{D^q}{q!} \sum_{j=2}^{n} j^{-r+q(1-\beta)} \\
&\leq Cn^{1-r} \sum_{q \geq q_\star} \frac{D^q}{q!} \frac{n^{q(1-\beta)}}{(q+1)(1-\beta)} \frac{(q+1)(1-\beta)}{q(1-\beta)+1-r} + C'n \\
&\leq Cn^{\beta-r} \sum_{q \geq q_\star} \frac{D^q}{(q+1)!} n^{(q+1)(1-\beta)} + C'n \\
&\leq C \exp(Dn^{1-\beta}) n^{\beta-r}.
\end{aligned}$$

This concludes the proof. $\square$

Lemma 11. *Let $\{\mathsf{A}_n, n \geq 0\}$ be a sequence of $d' \times q$ matrices and $\{\sigma_n, n \geq 0\}$ be a sequence of $q \times 1$ vectors.*

Let $\{S_n^{\mathrm{sa}}, n \geq 0\}$ *be given by* (12). *For any* $n \geq 2$

$$\begin{aligned}\mathsf{A}_n\left(S_n^{\mathrm{sa}} - \sigma_{n-1}\right) &= \Delta_{2:n}\mathsf{A}_1\left(S_1^{\mathrm{sa}} - \sigma_0\right) \\ &\quad + \sum_{j=2}^{n} \Delta_{j:n}\left(\mathsf{A}_j - \mathsf{A}_{j-1}\right)\left(S_{j-1}^{\mathrm{sa}} - \sigma_{j-2}\right) \\ &\quad + \sum_{j=2}^{n} \Delta_{j:n}\mathsf{A}_j\left(\sigma_{j-2} - \sigma_{j-1}\right) \\ &\quad + \sum_{j=2}^{n} \Delta_{j+1:n}\delta_j\mathsf{A}_j\left(m_j^{-1}\sum_{k=1}^{m_j} S(Z_{k,j-1}) - \sigma_{j-1}\right).\end{aligned}$$

Proof. By definition of S_n^{sa}, it holds $\mathsf{A}_n\left(S_n^{\mathrm{sa}} - \sigma_{n-1}\right) = (1-\delta_n)\mathsf{A}_{n-1}\left(S_{n-1}^{\mathrm{sa}} - \sigma_{n-2}\right) + B_n$ where

$$\begin{aligned}B_n &:= (1-\delta_n)(\mathsf{A}_n - \mathsf{A}_{n-1})\left(S_{n-1}^{\mathrm{sa}} - \sigma_{n-2}\right) \\ &\quad + (1-\delta_n)\mathsf{A}_n\left(\sigma_{n-2} - \sigma_{n-1}\right) \\ &\quad + \delta_n\mathsf{A}_n\left(m_n^{-1}\sum_{k=1}^{m_n} S(Z_{k,n-1}) - \sigma_{n-1}\right).\end{aligned}$$

By iterating, we have

$$\mathsf{A}_n\left(S_n^{\mathrm{sa}} - \sigma_{n-1}\right) = \Delta_{2:n}\mathsf{A}_1\left(S_1^{\mathrm{sa}} - \sigma_0\right) + \sum_{j=2}^{n}\Delta_{j+1:n}B_j,$$

from which the lemma follows. □

Lemma 12. *Assume H4a). Let* $\{S_n^{\mathrm{sa}}, n \geq 0\}$ *be given by* (12). *Then*

$$\begin{aligned}&\sup_{n\geq 0}\mathbb{E}\left[\|m_n^{-1}\sum_{j=1}^{m_n} S(Z_{j,n-1})\|^2\right] < \infty, \\ &\sup_{n\geq 0}\mathbb{E}\left[m_n^{-1}\sum_{j=1}^{m_n} W(Z_{j,n-1})\right] < \infty, \\ &\sup_{n\geq 0}\mathbb{E}\left[\|S_n^{\mathrm{sa}}\|^2\right] < \infty.\end{aligned}$$

Proof. By H4a), there exists a constant $C < \infty$ such that for any $n \geq 1$ and $1 \leq j \leq m_n$, $\|S(Z_{j,n-1})\|^2 \leq C\,W(Z_{j,n-1})$. In addition, by the drift assumption on the kernels P_θ, we have

$$\begin{aligned}\mathbb{E}\left[W(Z_{j,n-1})\right] &= \mathbb{E}\left[P_{\theta_{n-1}}W(Z_{j-1,n-1})\right] \\ &\leq \lambda\mathbb{E}\left[W(Z_{j-1,n-1})\right] + b \\ &\leq \lambda^j\mathbb{E}\left[W(Z_{0,n-1})\right] + \frac{b}{1-\lambda}.\end{aligned}$$

Similarly, by using $Z_{0,n-1} = Z_{m_{n-1},n-2}$, we have

$$\mathbb{E}\left[W(Z_{0,n-1})\right] \leq \lambda^{m_{n-1}}\mathbb{E}\left[W(Z_{0,n-2})\right] + \frac{b}{1-\lambda}.$$

A trivial induction shows that

$$\sup_n \sup_{j \le m_n} \mathbb{E}\left[W(Z_{j,n-1})\right] < \infty,$$

from which the first two results follow. For the third one: by Lemma 11 applied with $\mathsf{A}_n = I$ (the identity matrix) and $\sigma_n = 0$, we have for any $n \ge 1$,

$$S_n^{\mathrm{sa}} = \Delta_{2:n} S_1 + \sum_{j=2}^{n} \Delta_{j+1:n}\, \delta_j\, m_j^{-1} \sum_{k=1}^{m_j} S(Z_{k,j-1}).$$

By the Minkowsky inequality and the inequality $(a+b)^2 \le 2a^2 + 2b^2$, we have

$$\mathbb{E}\left[\|S_n^{\mathrm{sa}}\|^2\right] \quad \le \quad 2\left(\Delta_{2:n}\right)^2 \mathbb{E}\left[\|S_1^{\mathrm{sa}}\|^2\right] \quad + \quad 2 \sup_j \mathbb{E}\left[\|m_j^{-1} \sum_{k=1}^{m_j} S(Z_{k,j-1})\|^2\right] \left(\sum_{j=2}^{n} \Delta_{j+1:n}\, \delta_j\right)^2.$$

By definition, $\Delta_{2:n} \in [0,1]$ and by Lemma 8,

$$\sup_n \sum_{j=2}^{n} \Delta_{j+1:n}\, \delta_j < \infty.$$

Hence, $\sup_{n \ge 0} \mathbb{E}\left[\|S_n^{\mathrm{sa}}\|^2\right] < \infty$. □

Define the Proximal-Gradient operator

$$T_\gamma(\theta) := \mathrm{Prox}_{\gamma,g}\left(\theta + \gamma \nabla \ell(\theta)\right).$$

Lemma 13. *Assume H1, H2 and H3. Let $\{S_n^{\mathrm{sa}}, n \ge 0\}$ be given by (12). Then, for the sequence $\{\theta_n, n \ge 0\}$ given by Algorithm 2,*

(i) There exists a constant C such that almost-surely, for any $n \ge 0$,

$$\|\theta_{n+1} - \theta_n\| \le C \gamma_{n+1} \left(1 + \|S_{n+1}^{\mathrm{sa}} - \overline{S}(\theta_n)\|\right).$$

(ii) There exists a constant C' such that almost-surely, for any $n \ge 0$,

$$\|\gamma_{n+1}\Psi(\theta_n) \quad - \quad \gamma_n \Psi(\theta_{n-1})\| \quad \le \quad C'\left(|\gamma_{n+1} - \gamma_n| + \gamma_n^2 (1 + \|S_n^{\mathrm{sa}} - \overline{S}(\theta_{n-1})\|)\right).$$

(iii) There exists a constant C" such that almost-surely, for any $n \ge 0$,

$$\|\gamma_{n+1} T_{\gamma_{n+1},g}\left(\theta_n\right) \; - \; \gamma_n T_{\gamma_n,g}\left(\theta_{n-1}\right)\| \;\; \le \;\; C''\left(|\gamma_{n+1} - \gamma_n| + \gamma_n \gamma_{n+1} \; + \gamma_n^2 \left(1 + \|S_n^{\mathrm{sa}} - \overline{S}(\theta_{n-1})\|\right)\right).$$

Proof. The proof of *(i)* is on the same lines as the proof of [Atchadé et al., 2016, Lemma 15], and is omitted. For *(ii)*, we write by using H3b) and H3c),

$$\begin{aligned}
&\|\gamma_{n+1}\Psi(\theta_n) - \gamma_n \Psi(\theta_{n-1})\| \\
&\quad \le |\gamma_{n+1} - \gamma_n|\, \|\Psi(\theta_n)\| + \gamma_n \|\Psi(\theta_n) - \Psi(\theta_{n-1})\| \\
&\quad \le |\gamma_{n+1} - \gamma_n| \sup_\Theta \|\Psi\| + L\, \gamma_n \|\theta_n - \theta_{n-1}\|.
\end{aligned}$$

We then conclude by *(i)*. The LHS in *(iii)* is upper bounded by

$$|\gamma_{n+1}-\gamma_n| \sup_{\gamma\in(0,1/L]}\sup_{\theta\in\Theta}\|T_{\gamma,g}(\theta)\| + \gamma_n\|T_{\gamma_{n+1},g}(\theta_n)-T_{\gamma_n,g}(\theta_{n-1})\|.$$

Under H3, there exists a constant C such that for all $\gamma,\gamma'\in(0,1/L]$ and $\theta,\theta'\in\Theta$ (see [Atchadé et al., 2016, Proposition 12])

$$\sup_{\gamma\in(0,1/L]}\sup_{\theta\in\Theta}\|T_{\gamma,g}(\theta)\|<\infty,$$
$$\|T_{\gamma,g}(\theta)-T_{\gamma',g}(\theta')\|\le C\left(\gamma+\gamma'+\|\theta-\theta'\|\right).$$

We then conclude by *(i)*. □

Lemma 14. *Assume H4. For any $\theta\in\Theta$, there exists a function $\widehat{S}_\theta:\mathsf{Z}\to\mathbb{R}^q$ such that $S-\overline{S}(\theta)=\widehat{S}_\theta-P_\theta\widehat{S}_\theta$ and $\sup_{\theta\in\Theta}\left|\widehat{S}_\theta\right|_{\sqrt{W}}<\infty$. In addition, there exists a constant C such that for any $\theta,\theta'\in\Theta$,*

$$\left|P_\theta\widehat{S}_\theta-P_{\theta'}\widehat{S}_{\theta'}\right|_{\sqrt{W}}\le C\,\|\theta-\theta'\|.$$

Proof. Set $\widehat{S}_\theta(z):=\sum_{n\ge0}\left(P_\theta^nS(z)-\overline{S}(\theta)\right)$. Observe that, when exists, this function satisfies $S-\overline{S}(\theta)=\widehat{S}_\theta-P_\theta\widehat{S}_\theta$. Note that under H4a)-H4b), there exist C and $\rho\in(0,1)$ such that for any $\theta\in\Theta$,

$$\sum_{n\ge0}\left\|P_\theta^nS(z)-\overline{S}(\theta)\right\|\le C\,|S|_{\sqrt{W}}\left(\sum_{n\ge0}\rho^n\right)\sqrt{W}(z);$$

the RHS is finite, thus showing that $\widehat{S}_\theta$ exists. This inequality also proves that $\sup_\theta\left|\widehat{S}_\theta\right|_{\sqrt{W}}<\infty$. The Lipschitz property is established in [Fort et al., 2011a, Lemma 4.2.] and its proof uses H4c). □

5.3 Proof of Proposition 5

Throughout this section, set $\|U\|_{L_2}:=\mathbb{E}\left[\|U\|^2\right]^{1/2}$. By Lemma 11, $\|S_n^{\mathrm{sa}}-\overline{S}(\theta_{n-1})\|_{L_2}\le\sum_{i=1}^3\mathcal{T}_{i,n}$ with

$$\mathcal{T}_{1,n}:=\Delta_{2:n}\|S_1^{\mathrm{sa}}-\overline{S}(\theta_0)\|_{L_2},$$
$$\mathcal{T}_{2,n}:=\sum_{j=2}^n\Delta_{j:n}\|\overline{S}(\theta_{j-1})-\overline{S}(\theta_{j-2})\|_{L_2},$$
$$\mathcal{T}_{3,n}:=\left\|\sum_{j=2}^n\Delta_{j+1:n}\delta_j\left(m_j^{-1}\sum_{k=1}^{m_j}S(Z_{k,j-1})-\overline{S}(\theta_{j-1})\right)\right\|_{L_2}.$$

Since $\Delta_{2:n}\le\exp(-\delta_\star\sum_{j=2}^n j^{-\beta})$, then

$$\mathcal{T}_{1,n}=O\left(\exp(-\delta_\star(1-\beta)^{-1}n^{1-\beta})\right).$$

By H3c), Lemma 12 and Lemma 13, there exists a constant C such that $\mathcal{T}_{2,n}\le C\sum_{j=2}^n\Delta_{j:n}\gamma_{j-1}$. By Lemma 10, this yields $\mathcal{T}_{2,n}=O(n^{\beta-\alpha})$. For the last term, we use a martingale decomposition.

By Lemma 14, there exists a function $\widehat{S}_\theta$ such that

$$S(Z_{k,j-1}) - \overline{S}(\theta_{j-1}) = \widehat{S}_{\theta_{j-1}}(Z_{k,j-1}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{k,j-1}),$$

and $\sup_{\theta\in\Theta}\left|\widehat{S}_\theta\right|_{\sqrt{W}} < \infty$. Hence, we write

$$m_j^{-1}\sum_{k=1}^{m_j} S(Z_{k,j-1}) - \overline{S}(\theta_{j-1}) = \partial M_j + R_{j,1} + R_{j,2}$$

with

$$\begin{aligned}
\partial M_j &:= m_j^{-1}\sum_{k=1}^{m_j}\left\{\widehat{S}_{\theta_{j-1}}(Z_{k,j-1}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{k-1,j-1})\right\},\\
R_{j,1} &:= m_j^{-1}\left\{P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{0,j-1}) - P_{\theta_j}\widehat{S}_{\theta_j}(Z_{0,j})\right\},\\
R_{j,2} &:= m_j^{-1}\left\{P_{\theta_j}\widehat{S}_{\theta_j}(Z_{0,j}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{0,j})\right\};
\end{aligned}$$

we used that $Z_{0,j} = Z_{m_j,j-1}$. Upon noting that ∂M_j is a martingale-increment, and

$$\widehat{S}_{\theta_{j-1}}(Z_{k,j-1}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{k-1,j-1})$$

is a martingale-increment, we have by two successive applications of [Hall and Heyde, 1980, Theorem 2.10]:

$$\left\|\sum_{j=2}^{n}\Delta_{j+1:n}\delta_j\partial M_j\right\|_{L_2} \le C\left(\sum_{j=2}^{n}\Delta_{j+1:n}^2\frac{\delta_j^2}{m_j}\right)^{1/2}.$$

By Lemma 10, this term is $O(n^{-(\beta+c)/2})$. For the second term, we write

$$\sum_{j=2}^{n}\Delta_{j+1:n}\delta_j R_{j,1} = \Delta_{3:n}\frac{\delta_2}{m_2}P_{\theta_1}\widehat{S}_{\theta_1}(Z_{0,1}) - \frac{\delta_n}{m_n}P_{\theta_n}\widehat{S}_{\theta_n}(Z_{0,n}) + \sum_{j=2}^{n-1}\left(\Delta_{j+2:n}\frac{\delta_{j+1}}{m_{j+1}} - \Delta_{j+1:n}\frac{\delta_j}{m_j}\right)P_{\theta_j}\widehat{S}_{\theta_j}(Z_{0,j}).$$

By Lemma 12 and Lemma 14, the RHS is $O(n^{-(\beta+c)} + n^{-(1+c)})$ so that this second term is $O(n^{-(\beta+c)})$. Finally, for the third term, by using Lemma 12, Lemma 13 and Lemma 14, we write

$$\left\|\sum_{j=2}^{n}\Delta_{j+1:n}\delta_j R_{j,2}\right\|_{L_2} \le \sum_{j=2}^{n}\Delta_{j+1:n}\frac{\delta_j}{m_j}\gamma_j.$$

Again by Lemma 10, this last term is $O(n^{-(\alpha+c)})$. Therefore, $\mathcal{T}_{3,n} = O(n^{-(\beta+c)/2})$.

5.4 Proof of Theorem 6

Throughout the proof, we will write S_{n+1} instead of S_{n+1}^{sa}.

Proof of Theorem 6 We prove the almost-sure convergence of the three random sums given in Theorem 2. The third one is finite almost-surely since its expectation is finite (see Proposition 15). The first two ones are of the form $\sum_n \mathsf{A}_{n+1}\left(S_{n+1}-\overline{S}(\theta_n)\right)$ where A_{n+1} is respectively

$$\mathsf{A}_{n+1}=\gamma_{n+1}\left(T_{\gamma_{n+1},g}\left(\theta_n\right)\right)', \qquad \mathsf{A}_{n+1}=\gamma_{n+1}\Psi(\theta_n).$$

Note that $\mathsf{A}_{n+1}\in\mathcal{F}_n$ (the filtration is defined by (29)). By Lemma 13 and H3b-c), for both cases, there exists a constant C such that almost-surely, for any $n\geq 0$,

$$\begin{aligned}\|\mathsf{A}_{n+1}-\mathsf{A}_n\| &\leq C\left(|\gamma_{n+1}-\gamma_n|+\gamma_n^2+\gamma_n\gamma_{n+1}\right)\ \cdots\\ &\qquad\times\left(1+\|S_n-\overline{S}(\theta_{n-1})\|\right),\\ \|\mathsf{A}_{n+1}\| &\leq C\gamma_{n+1}.\end{aligned}$$

We then conclude by Proposition 16.

Proposition 15. *Assume H4a) and*

$$\sup_{\theta\in\Theta}\left(\|\Psi(\theta)\|+\|\overline{S}(\theta)\|\right)<\infty.$$

Then there exists a constant C such that

$$\sum_n \gamma_{n+1}^2\mathbb{E}\left[\left\|\Psi(\theta_n)\left(S_{n+1}-\overline{S}(\theta_n)\right)\right\|^2\right]\leq C\sum_n\gamma_{n+1}^2.$$

Proof. We write

$$\mathbb{E}\left[\left\|\Psi(\theta_n)\left(S_{n+1}-\overline{S}(\theta_n)\right)\right\|^2\right] \qquad\leq\qquad 2\sup_{\Theta}\|\Psi(\theta)\|^2\left(\sup_n\mathbb{E}\left[\|S_n\|^2\right]+\sup_{\Theta}\|\overline{S}\|^2\right),$$

and conclude by Lemma 12. □

Proposition 16. *Let $\{\theta_n, n\geq 0\}$ be given by Algorithm 2. Assume H1, H3, H4a-b) and H5a). In the biased case, assume also H4c) and H5b). Let $\{\mathsf{A}_n, n\geq 0\}$ be a sequence of $d'\times q$ random matrices such that for any $n\geq 0$, $\mathsf{A}_{n+1}\in\mathcal{F}_n$, and there exists a constant $C_\star$ such that almost-surely*

$$\|\mathsf{A}_{n+1}\|\leq C_\star\gamma_{n+1}, \tag{30}$$

$$\|\mathsf{A}_{n+1}-\mathsf{A}_n\|\leq C_\star a_{n+1}\left(1+\|S_n-\overline{S}(\theta_{n-1})\|\right); \tag{31}$$

here $a_{n+1}:=\gamma_n\gamma_{n+1}+\gamma_n^2+|\gamma_{n+1}-\gamma_n|$. Then, almost-surely, the series $\sum_n\mathsf{A}_{n+1}\left(S_{n+1}-\overline{S}(\theta_n)\right)$ converges.

By Lemma 11 applied with $\sigma_n = \overline{S}(\theta_n)$, we decompose this sum into four terms:

$$\begin{aligned}
\mathcal{T}_1 &:= \sum_{n\geq 2} \Delta_{2:n}\mathsf{A}_1 \left(S_1 - \overline{S}(\theta_0)\right) \\
&= \mathsf{D}_2\mathsf{A}_1 \left(S_1 - \overline{S}(\theta_0)\right), \\
\mathcal{T}_2 &:= \sum_{n\geq 2}\sum_{j=2}^{n} \Delta_{j:n} \left(\mathsf{A}_j - \mathsf{A}_{j-1}\right) \left(S_{j-1} - \overline{S}(\theta_{j-2})\right) \\
&= \sum_{j\geq 2} \mathsf{D}_j \left(\mathsf{A}_j - \mathsf{A}_{j-1}\right) \left(S_{j-1} - \overline{S}(\theta_{j-2})\right), \\
\mathcal{T}_3 &:= \sum_{n\geq 2}\sum_{j=2}^{n} \Delta_{j:n}\mathsf{A}_j \left(\overline{S}(\theta_{j-2}) - \overline{S}(\theta_{j-1})\right) \\
&= \sum_{j\geq 2} \mathsf{D}_j\mathsf{A}_j \left(\overline{S}(\theta_{j-2}) - \overline{S}(\theta_{j-1})\right), \\
\mathcal{T}_4 &:= \sum_{n\geq 2}\sum_{j=2}^{n} \delta_j\Delta_{j+1:n}\mathsf{A}_j \left(m_j^{-1}\sum_{k=1}^{m_j} S(Z_{k,j-1}) - \overline{S}(\theta_{j-1})\right) \\
&= \sum_{j\geq 2} \delta_j(1+\mathsf{D}_{j+1})\mathsf{A}_j \left(m_j^{-1}\sum_{k=1}^{m_j} S(Z_{k,j-1}) - \overline{S}(\theta_{j-1})\right).
\end{aligned}$$

We have by using (30),

$$\mathsf{D}_2\|\mathsf{A}_1 \left(S_1 - \overline{S}(\theta_0)\right)\| \leq \mathsf{D}_2 C_\star\gamma_1 \left(\|S_1\| + \sup_{\Theta}\|\overline{S}\|\right).$$

By H5a), $\mathsf{D}_2 < \infty$ so the RHS is finite thus implying that $\mathcal{T}_1$ is finite almost-surely.
Using (31), there exists a constant C such that

$$\mathbb{E}\left[\sum_{j\geq 2}\mathsf{D}_j\|\left(\mathsf{A}_j - \mathsf{A}_{j-1}\right)\left(S_{j-1} - \overline{S}(\theta_{j-2})\right)\|\right] \leq C\left(1 + \sup_n \mathbb{E}\left[\|S_n\|^2\right] + \sup_{\Theta}\|\overline{S}\|\right)\sum_{j\geq 2} a_j\mathsf{D}_j.$$

By H3b)-H3c), H5a) and Lemma 12, the RHS is finite thus implying that $\mathcal{T}_2$ is finite almost-surely.
Similarly, there exists a constant C such that

$$\mathbb{E}\left[\sum_{j\geq 2}\mathsf{D}_j\|\mathsf{A}_j\left(\overline{S}(\theta_{j-2}) - \overline{S}(\theta_{j-1})\right)\|\right] \leq C\sum_{j\geq 2}\gamma_j\mathsf{D}_j\mathbb{E}\left[\|\overline{S}(\theta_{j-1}) - \overline{S}(\theta_{j-2})\|\right].$$

By H3c), the RHS is bounded (up to a multiplicative constant) by $\sum_{j\geq 2}\gamma_j\mathsf{D}_j\mathbb{E}\left[\|\theta_{j-1} - \theta_{j-2}\|\right]$; and by H5a) and Lemmas 12 and 13, this sum is finite. Hence $\mathcal{T}_3$ is finite almost-surely.
We give the proof of the convergence of the last term in the biased case: $\mathbb{E}\left[S(Z_{k,n})|\mathcal{F}_n\right] \neq \overline{S}(\theta_n)$. The proof in the unbiased case corresponds to the following lines with $R_{j,1} = R_{j,2} = 0$ and $\widehat{S}_\theta = S$. Set $\overline{\mathsf{D}}_j := \delta_j(1+\mathsf{D}_{j+1})$.
By Lemma 14, there exists $\widehat{S}_\theta$ such that

$$S(Z_{k,j-1}) - \overline{S}(\theta_{j-1}) = \widehat{S}_{\theta_{j-1}}(Z_{k,j-1}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{k,j-1}),$$

and $\sup_{\theta\in\Theta}\left|\widehat{S}_\theta\right|_{\sqrt{W}} < \infty$. Hence, we have

$$\mathcal{T}_4 = \sum_{j\geq 2} \overline{\mathsf{D}}_j \mathsf{A}_j \left(\partial M_j + R_{j,1} + R_{j,2}\right),$$

where

$$\begin{aligned}
\partial M_j &:= m_j^{-1} \sum_{k=1}^{m_j} \left(\widehat{S}_{\theta_{j-1}}(Z_{k,j-1}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{k-1,j-1})\right), \\
R_{j,1} &:= m_j^{-1}\left(P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{0,j-1}) - P_{\theta_j}\widehat{S}_{\theta_j}(Z_{0,j})\right), \\
R_{j,2} &:= m_j^{-1}\left(P_{\theta_j}\widehat{S}_{\theta_j}(Z_{0,j}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{0,j})\right).
\end{aligned}$$

Upon noting that $\mathbb{E}\left[\mathsf{A}_j \partial M_j | \mathcal{F}_{j-1}\right] = 0$, the almost-sure convergence of the series $\sum_j \overline{\mathsf{D}}_j \mathsf{A}_j \partial M_j$ is proved by checking criteria for the almost-sure convergence of a martingale. By (30), there exists a constant C such that

$$\sum_j \overline{\mathsf{D}}_j^2 \mathbb{E}\left[\|\mathsf{A}_j \partial M_j\|^2\right] \quad \leq \quad C \sum_j \frac{\gamma_j^2}{m_j^2}\overline{\mathsf{D}}_j^2 \quad \times \quad \mathbb{E}\left[\left\|\sum_{k=1}^{m_j}\left(\widehat{S}_{\theta_{j-1}}(Z_{k,j-1}) - P_{\theta_{j-1}}\widehat{S}_{\theta_{j-1}}(Z_{k-1,j-1})\right)\right\|^2\right].$$

By H5a), Lemma 12 and [Hall and Heyde, 1980, Theorem 2.10], the RHS is finite. [Hall and Heyde, 1980, Theorem 2.17] implies that $\sum_j \overline{\mathsf{D}}_j \mathsf{A}_j\, \partial M_j$ is finite almost-surely. For the second term, we write

$$\sum_{j\geq 2} \overline{\mathsf{D}}_j \mathsf{A}_j R_{j,1} \quad = \quad m_2^{-1}\overline{\mathsf{D}}_2 \mathsf{A}_2 P_{\theta_1}\widehat{S}_{\theta_1}(Z_{0,1}) \quad + \quad \sum_{j\geq 2}\left(m_{j+1}^{-1}\overline{\mathsf{D}}_{j+1}\mathsf{A}_{j+1} - m_j^{-1}\overline{\mathsf{D}}_j\mathsf{A}_j\right) P_{\theta_j}\widehat{S}_{\theta_j}(Z_{0,j}),$$

so that, by Lemmas 12 and 14, this series is finite almost-surely if $\sum_j \mathbb{E}\left[\|m_{j+1}^{-1}\overline{\mathsf{D}}_{j+1}\mathsf{A}_{j+1} - m_j^{-1}\overline{\mathsf{D}}_j\mathsf{A}_j\|\right] < \infty$. From (30) and (31), there exists a constant C such that

$$\|m_{j+1}^{-1}\overline{\mathsf{D}}_{j+1}\mathsf{A}_{j+1} \;-\; m_j^{-1}\overline{\mathsf{D}}_j\mathsf{A}_j\| \;\leq\; C\gamma_{j+1}\left|m_{j+1}^{-1}\overline{\mathsf{D}}_{j+1} - m_j^{-1}\overline{\mathsf{D}}_j\right| \;+\; m_j^{-1}\overline{\mathsf{D}}_j a_{j+1}\left(1 + \|S_j - \overline{S}(\theta_{j-1})\|\right).$$

H5 and Lemma 12 imply that

$$\sum_j \mathbb{E}\left[\|m_{j+1}^{-1}\overline{\mathsf{D}}_{j+1}\mathsf{A}_{j+1} - m_j^{-1}\overline{\mathsf{D}}_j\mathsf{A}_j\|\right] < \infty.$$

Finally, by (30), Lemmas 12 to 14, there exists a constant C such that

$$\sum_{j\geq 2}\mathbb{E}\left[\overline{\mathsf{D}}_j\,\|\mathsf{A}_j R_{j,2}\|\right] \leq C\sum_{j\geq 2}\gamma_j^2 m_j^{-1}\overline{\mathsf{D}}_j.$$

The RHS is finite by H5 thus implying that $\sum_j \overline{\mathsf{D}}_j \mathsf{A}_j R_{j,2}$ is finite almost-surely.

References

C. Andrieu and E. Moulines. On the ergodicity properties of some adaptive MCMC algorithms. *Ann. Appl. Probab.*, 16(3):1462–1505, 2006.

C. Andrieu and J. Thoms. A tutorial on adaptive MCMC. *Statistics and Computing*, 18(4):343–373, 2008.

Y. Atchadé, G. Fort, and E. Moulines. On perturbed proximal gradient algorithms. *Accepted for publication in JMLR*, 2016.

D. Bartholomew, M. Knott, and I. Moustaki. *Latent variable models and factor analysis*. Wiley Series in Probability and Statistics. John Wiley & Sons, Ltd., Chichester, third edition, 2011. A unified approach.

H.H. Bauschke and P.L. Combettes. *Convex analysis and monotone operator theory in Hilbert spaces*. CMS Books in Mathematics/Ouvrages de Mathématiques de la SMC. Springer, New York, 2011.

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

A. Benveniste, M. Métivier, and P. Priouret. *Adaptive algorithms and stochastic approximations*, volume 22 of *Applications of Mathematics (New York)*. Springer-Verlag, Berlin, 1990. Translated from the French by Stephen S. Wilson.

J. Bertrand and D.J. Balding. Multiple single nucleotide polymorphism analysis using penalized regression in nonlinear mixed-effect pharmacokinetic models. *Pharmacogenetics and genomics*, 23(3):167–174, 2013.

P. J. Bickel and K. A. Doksum. *Mathematical statistics—basic ideas and selected topics. Vol. 1*. Texts in Statistical Science Series. CRC Press, Boca Raton, FL, second edition, 2015.

H. Chen, D. Zeng, and Y. Wang. Penalized nonlinear mixed effects model to identify biomarkers that predict disease progression. *Biometrics*, 2017.

J. Chen and Z. Chen. Extended Bayesian information criteria for model selection with large model spaces. *Biometrika*, 95(3):759–771, 2008.

P. L. Combettes and J.C. Pesquet. Stochastic quasi-Fejér block-coordinate fixed point iterations with random sweeping. *SIAM J. Optim.*, 25(2):1221–1248, 2015.

P. L. Combettes and V.R. Wajs. Signal recovery by proximal forward-backward splitting. *Multiscale Model. Simul.*, 4(4):1168–1200 (electronic), 2005.

P.L. Combettes and J.C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-point algorithms for inverse problems in science and engineering*, volume 49 of *Springer Optim. Appl.*, pages 185–212. Springer, New York, 2011.

P.L. Combettes and J.C. Pesquet. Stochastic Approximations and Perturbations in Forward-Backward Splitting for Monotone Operators. *Online journal Pure and Applied Functional Analysis*, 1(1):1–37, 2016.

X. Delavenne, E. Ollier, T. Basset, L. Bertoletti, S. Accassat, A. Garcin, S. Laporte, P. Zufferey, and P. Mismetti. A semi-mechanistic absorption model to evaluate drug–drug interaction with dabigatran: application with clarithromycin. *British journal of clinical pharmacology*, 76(1):107–113, 2013.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 1999.

A. P. Dempster, N. M. Laird, and D. B. Rubin. Maximum likelihood from incomplete data via the EM algorithm. *J. Roy. Statist. Soc. Ser. B*, 39(1):1–38, 1977.

G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6):3262–3289, 2011a.

G. Fort, E. Moulines, and P. Priouret. Convergence of adaptive and interacting Markov chain Monte Carlo algorithms. *Ann. Statist.*, 39(6), 2011b.

G. Fort, B. Jourdain, E. Kuhn, T. Lelièvre, and G. Stoltz. Convergence of the Wang-Landau algorithm. *Math. Comp.*, 84(295):2297–2327, 2015.

J. Friedman, T Hastie, and R. Tibshirani. Regularization paths for generalized linear models via coordinate descent. *Journal of Statistical Software*, 33(1):1–22, 2010. URL `http://www.jstatsoft.org/v33/i01/`.

I Gouin-Thibault, X Delavenne, A Blanchard, V Siguret, JE Salem, C Narjoz, P Gaussem, P Beaune, C Funck-Brentano, M Azizi, et al. Inter-individual variability in dabigatran and rivaroxaban exposure: contribution of abcb1 genetic polymorphisms and interaction with clarithromycin. *Journal of Thrombosis and Haemostasis*, 2016.

P. Hall and C. C. Heyde. *Martingale limit theory and its application*. Academic Press, Inc. [Harcourt Brace Jovanovich, Publishers], New York-London, 1980. Probability and Mathematical Statistics.

H.J. Kushner and G.G. Yin. *Stochastic approximation and recursive algorithms and applications*, volume 35 of *Applications of Mathematics (New York)*. Springer-Verlag, New York, second edition, 2003. Stochastic Modelling and Applied Probability.

E.L. Lehmann and G. Casella. *Theory of point estimation*. Springer Science & Business Media, 2006.

R. A. Levine and J. Fan. An automated (Markov chain) Monte Carlo EM algorithm. *J. Stat. Comput. Simul.*, 74(5):349–359, 2004. ISSN 0094-9655. doi: 10.1080/0094965031000147704. URL `http://dx.doi.org/10.1080/0094965031000147704`.

J. Lin, L. Rosasco, S. Villa, and D.X. Zhou. Modified Fejer Sequences and Applications. Technical report, arXiv:1510:04641v1 math.OC, 2015.

G.J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley Series in Probability and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.

R. R. Meyer. Sufficient conditions for the convergence of monotonic mathematical programming algorithms. *J. Comput. System. Sci.*, 12(1):108–121, 1976.

S. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. With a prologue by Peter W. Glynn.

J.J. Moreau. Fonctions convexes duales et points proximaux dans un espace hilbertien. *C. R. Acad. Sci. Paris*, 255:2897–2899, 1962.

S.K. Ng, T. Krishnan, and G.J. McLachlan. The EM algorithm. In *Handbook of computational statistics—concepts and methods. 1, 2*, Springer Handb. Comput. Stat., pages 139–172. Springer, Heidelberg, 2012.

E. Ollier, S. Hodin, T. Basset, S. Accassat, L. Bertoletti, P. Mismetti, and X. Delavenne. In vitro and in vivo evaluation of drug–drug interaction between dabigatran and proton pump inhibitors. *Fundamental & clinical pharmacology*, 29(6):604–614, 2015.

E. Ollier, A. Samson, X. Delavenne, and V. Viallon. A saem algorithm for fused lasso penalized nonlinear mixed effect models: Application to group comparison in pharmacokinetics. *Computational Statistics & Data Analysis*, 95:207–221, 2016.

N. Parikh and S. Boyd. Proximal Algorithms. *Foundations and Trends in Optimization*, 1(3):123–231, 2013.

C. P. Robert and G. Casella. *Monte Carlo statistical methods*. Springer Texts in Statistics. Springer-Verlag, New York, second edition, 2004.

G.O. Roberts and J.S. Rosenthal. Coupling and Ergodicity of Adaptive MCMC. *J. Appl. Probab.*, 44(2): 458–475, 2007.

L. Rosasco, S. Villa, and B.C. Vu. Convergence of a Stochastic Proximal Gradient Algorithm. Technical report, arXiv:1403.5075v3, 2014.

L. Rosasco, S. Villa, and B.C. Vu. A stochastic inertial forward–backward splitting algorithm for multivariate monotone inclusions. *Optimization*, 65(6):1293–1314, 2016.

E. Saksman and M. Vihola. On the ergodicity of the adaptive Metropolis algorithm on unbounded domains. *Ann. Appl. Probab.*, 20(6):2178–2203, 2010.

A. Samson, M. Lavielle, and F. Mentré. The SAEM algorithm for group comparison tests in longitudinal data analysis based on non-linear mixed-effects model. *Statistics in medicine*, 26(27):4860–4875, 2007.

A. Schreck, G. Fort, and E. Moulines. Adaptive equi-energy sampler: convergence and illustration. *ACM Trans. Model. Comput. Simul.*, 23(1):Art. 5, 27, 2013.

N. Städler, P. Bühlmann, and S. van de Geer. l1-penalization for mixture regression models. *TEST*, 19(2): 209–256, 2010.

G. Wei and M. Tanner. A Monte-Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *J. Amer. Statist. Assoc.*, 85:699–704, 1990.

C.-F.J. Wu. On the convergence properties of the EM algorithm. *Ann. Statist.*, 11(1):95–103, 1983.

W.I. Zangwill. *Nonlinear programming: a unified approach*. Prentice-Hall, Inc., Englewood Cliffs, N.J., 1969. Prentice-Hall International Series in Management.