MISSO: Minimization by Incremental Stochastic Surrogate for large-scale nonconvex Optimization

Belhal Karimi
Center of Applied Mathematics
Ecole Polytechnique
Palaiseau, FR
belhal.karimi@polytechnique.edu

Eric Moulines
Center of Applied Mathematics
Ecole Polytechnique
Palaiseau, FR
eric.moulines@polytechnique.edu

Abstract

Many nonconvex optimization problems can be solved using the Majorization-Minimization (MM) algorithm that consists in upper bounding, at each iteration of the algorithm, the objective function by a surrogate that is easier to minimize. When the objective function can be expressed as a large sum of individual losses, incremental version of the MM algorithm is often used. However, in many cases of interest (Generalized Linear Mixed Model or Variational Bayesian inference) those surrogates are intractable. In this contribution, we propose a generalization of incremental MM algorithm using Monte Carlo approximation of these surrogates. We establish the convergence of our unifying scheme for possibly nonconvex objective. Finally, we apply our new framework to train a logistic regression and a Bayesian neural network on the MNIST dataset and compare its convergence behaviour with state-of-the-art optimization methods.

1 Introduction

We are interested in the constrained minimization of a large sum of nonconvex functions defined as:

$$\min_{\theta \in \Theta} \left[f(\theta) \triangleq \sum_{i=1}^{N} f_i(\theta) \right] \tag{1}$$

where Θ is a convex subset of $\mathbb{R}^p$, for all $i \in [\![N]\!]$, $f_i : \mathbb{R}^p \to \mathbb{R}$ are continuously differentiable, bounded from below and possibly nonconvex. In this paper, we solve this minimization problem using an MM algorithm [Lange, 2016, Razaviyayn et al., 2013] which works by finding iteratively a surrogate function that majorizes the objective function. By minimizing at each iteration the surrogate function, we drive the objective function downwards until convergence to a stationary point. MM algorithms become very popular in machine learning and computational statistics [Lange, 2016]. Examples include the proximal gradient algorithm [Beck and Teboulle, 2009, Parikh and Boyd, 2014], the Expectation-Maximization (EM) algorithm [McLachlan and Krishnan, 2007] and some variational inference methods [Wainwright and Jordan, 2008].

When the objective function is a finite-sum, [Mairal, 2015] developed an incremental MM scheme, called MISO, taking advantage of the finite-sum structure with a cost per iteration that is independent of N. Incremental methods have recently become very popular; in particular these methods proved to be an essential component to develop variance reduced stochastic gradient methods [Le Roux et al., 2012, Defazio et al., 2014]. See [Mairal, 2015] and the references therein.

However, the MISO framework rests upon the computation of tractable surrogates such as quadratic or variational functions. Yet, in many cases, those surrogates are intractable and need to be approximated. For instance, in the Bayesian machine learning literature [Ghahramani, 2015], uncertainty is put

on the parameters, which optimization problem boils down to finding the true distribution of those parameters given any observed data. To this end, variational inference methods, as approximate inference methods, [Ranganath et al., 2014, Kingma and Welling, 2013] have been extensively studied to find an approximation of this distribution which can also be used as a proposal for an exact sampler [Girolami and Calderhead, 2011, de Freitas et al., 2001]. More recently, Bayesian neural networks [Neal, 2012, I. Goodfellow and Courville, 2016], vastly studied in [Gal, 2016, Blundell et al., 2015, Mullachery et al., 2018, Polson and Sokolov, 2017, D. J. Rezende and Wierstra, 2014], can produce probabilistic guarantees on their predictions and also generate the distribution of the parameters that it has learnt from the observations. These two characteristics make them highly attractive to theoreticians as well as practitioners. Variational inference methods mentioned above, are extensively used [B. Trippe, 2018, Pawlowski et al., 2017, Y. Li, 2017] for training such neural network. To scale to large datasets, this optimization is typically performed using Stochastic Gradient Descent (SGD), one of its variants [Bottou et al., 2016] or using the Stochastic Variational Inference algorithm proposed in [Hoffman et al., 2013], and its variants [Kucukelbir et al., 2017, Titsias and Lázaro-Gredilla, 2014, Kingma and Welling, 2013] which approximates the full gradient from mini-batches. Ultimately, MISO convergence guarantees can not be applied on those cases where approximation of surrogates are used and they often rely on Robbins and Monro [Robbins and Monro, 1951] convergence results for stochastic optimization.

In Generalized Linear Mixed Models, Maximum Likelihood Estimation is performed to fit the parameters of a model to the observed data. Random effects are considered as latent variables and the optimization procedure requires augmenting the observed data with the latent structure. The EM algorithm [McLachlan and Krishnan, 2007] is a reference method to execute this task. In particular, the Incremental EM, introduced by [Neal and Hinton, 1998], updates upper-bounds of the negated log-likelihood incrementally and can be shown to be a special case of the MISO framework. When those upper-bounds are intractable, the MCEM [Wei and Tanner, 1990] algorithm optimizes their Monte Carlo integrations. While many convergence results of this algorithm have been provided [Fort and Moulines, 2003, Neath, 2012], its mini-batch version is not guaranteed to converge.

In this contribution, we propose an incremental MM algorithm, called MISSO (Minimization by Incremental Stochastic Surrogate Optimization) when the natural surrogate functions are intractable and should be approximated, for example by Monte Carlo integration. We present a unifying framework in which the mini-batch MCEM and the mini-batch Variational Inference algorithm, an extension of the Stochastic Variational Inference that incorporates a memory of previous gradients, fall under and provide convergence guarantees of the objective function. Finally, we apply our incremental MM scheme to train a logistic regression on synthetic data and a Bayesian neural network on MNIST dataset [LeCun and Cortes, 2010] to highlight the effectiveness of our method.

2 Incremental minimization of large sum of nonconvex functions

Beforehand, let $\mathcal{T}(\Theta)$ be a neighborhood of Θ and assume that:

M 1. *For all $i \in [\![N]\!]$, f_i is continuously differentiable on $\mathcal{T}(\Theta)$.*

M 2. *For all $i \in [\![N]\!]$, f_i is bounded from below, i.e. there exist a constant $M_i \in \mathbb{R}$ such as for all $\theta \in \Theta$, $f_i(\theta) \geq M_i$.*

For any $\theta \in \Theta$ and $i \in [\![N]\!]$, we say, following [Mairal, 2015] that a function $f_{i,\theta} : \mathbb{R}^p \to \mathbb{R}$ is a surrogate of f_i at θ if the following properties are satisfied:

S.1 the function $\vartheta \to f_{i,\theta}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$

S.2 for all $\vartheta \in \Theta$, $f_{i,\theta}(\vartheta) \geq f_i(\vartheta)$, $f_{i,\theta}(\theta) = f_i(\theta)$ and $\nabla f_{i,\theta}(\vartheta)\big|_{\vartheta=\theta} = \nabla f_i(\vartheta)\big|_{\vartheta=\theta}$.

The gap $f_{i,\theta} - f_i$ plays a key role in the convergence analysis and we require this error to be L-smooth for some constant $L > 0$ in the following sense:

Definition 1. (*L-smooth functions*) A function $f : \Theta \to \mathbb{R}$ is called *L*-smooth when it is differentiable and when its gradient ∇f is *L*-Lipschitz continuous.

Denote by $\langle \cdot, \cdot \rangle$ the scalar product, we also introduce the following stationary point condition:

Definition 2. (Asymptotic Stationary Point Condition)
A sequence $(\theta^k)_{k \geq 0}$ satisfies the asymptotic stationary point condition if

$$\liminf_{k \to \infty} \inf_{\theta \in \Theta} \frac{\langle \nabla f(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq 0. \tag{2}$$

The incremental scheme of [Mairal, 2015] computes surrogate functions, at each iteration of the algorithm, for a mini-batch of components:

Algorithm 1 MISO algorithm

Initialization: given an initial parameter estimate θ^0, for all $i \in [\![N]\!]$ compute a surrogate function $\vartheta \to f_{i,\theta^0}(\vartheta)$.
Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![N]\!], \text{card}(A) = p\}$

2. For all $i \in I_k$ and compute $\vartheta \to f_{i,\theta^{k-1}}(\vartheta)$, a surrogate of f_i at θ^{k-1}.

3. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^{N} a_i^k(\vartheta)$ where $a_i^k(\vartheta)$ are defined recursively as follows:

$$a_i^k(\vartheta) \triangleq \begin{cases} f_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ a_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{3}$$

For all $i \in [\![N]\!]$ and $\vartheta \in \Theta$, $a_i^k(\vartheta) = f_{i,\theta^{\tau_{i,k}}}(\vartheta)$ where for all $i \in [\![N]\!]$, $\tau_{i,0} = 0$ and $k \geq 1$ the indices $\tau_{i,k}$ are defined recursively as follows:

$$\tau_{i,k} = k - 1 \quad \text{if } i \in I_k \quad \text{and } \tau_{i,k} = \tau_{i,k-1} \text{ otherwise} \tag{4}$$

Let $(\theta^k)_{k \geq 1}$ be a sequence generated from $\theta^0 \in \Theta$ by the iterative application described by Algorithm 1 then, in [Mairal, 2015], almost sure convergence of the sequence $(f(\theta^k))_{k \geq 1}$ is established and $(\theta^k)_{k \geq 1}$ is shown to satisfy the Asymptotic Stationary Point Condition.

2.1 Minimization by Incremental Stochastic Surrogate Optimization (MISSO) scheme

In this section, we introduce an incremental scheme when the surrogate functions computed in Algorithm 1 are not tractable. We assume that the surrogate can be expressed as an integral over a set of latent variables, denoted $z = (z_i \in Z_i, i \in [\![N]\!]) \in Z$ where $Z = \bigtimes_{i=1}^{N} Z_i$ where Z_i is a subset of $\mathbb{R}^{m_i}$. For all $i \in [\![N]\!]$, let μ_i be a σ-finite measure on the Borel σ-algebra $\mathcal{Z}_i = \mathcal{B}(Z_i)$, $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ be a family of probability densities with respect to μ_i, and $r_i : Z_i \times \Theta \to \mathbb{R}$ be functions such that:

$$f_{i,\theta}(\vartheta) \triangleq \int_{Z_i} r_{i,\theta}(z_i, \vartheta) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \quad \text{for all } (\theta, \vartheta) \in \Theta^2. \tag{5}$$

The surrogate function denoted $f_{i,\theta}(\vartheta)$ is fully defined by the pair $(r_{i,\theta}(z_i, \vartheta), p_i(z_i, \theta))$.

Example (Incremental EM). *The Expectation-Maximization (EM) algorithm is the reference method to perform Maximum Likelihood Estimation in incomplete data problem [McLachlan and Krishnan, 2007]. Let $\{c_i(z_i, \theta), \theta \in \Theta\}$ be a family of positive μ_i-integrable Borel functions on Z_i. Define, for all $i \in [\![N]\!]$ and $\theta \in \Theta$, $g_i(\theta) \triangleq \int_{Z_i} c_i(z_i, \theta) \mu_i(\mathrm{d}z_i)$. In the missing data context, $c_i(z_i, \theta)$ is the joint likelihood of the observations and the latent data referred to as the complete likelihood and $g_i(\theta)$ is the likelihood of the observations (in which the latent variables are marginalized). The incremental EM algorithm falls into the incremental MM framework outlined above. In such case, for $i \in [\![N]\!]$ and $\theta \in \Theta$ the loss function $f_i(\theta)$ is the negated incomplete log-likelihood $f_i(\theta) \triangleq -\log g_i(\theta)$, for $\vartheta \in \Theta$ the surrogate function $f_{i,\theta}(\vartheta)$ is defined by the pair $(r_{i,\theta}(z_i, \vartheta), p_i(z_i, \theta))$ such as:*

$$r_{i,\theta}(z_i, \vartheta) \triangleq \log \left(p_i(z_i, \theta) / c_i(z_i, \vartheta) \right) \quad \text{and} \quad p_i(z_i, \theta) \triangleq c_i(z_i, \theta) / g_i(\theta) \quad \text{if } g_i(\theta) \neq 0 \tag{6}$$

With these notations, the MISO algorithm outlined in Algorithm 1 coincides with the incremental EM algorithm introduced in the pioneering paper [Neal and Hinton, 1998] by Neal and Hinton.

3

Example (Incremental Variational Inference for latent data model). *Let $x = (x_i, i \in [\![N]\!])$ and $y = (y_i, i \in [\![N]\!])$ be i.i.d. input-output pairs and w be a global latent variable taking values in W as subset of $\mathbb{R}^J$. A natural decomposition of the joint distribution is:*

$$p(y, x, w) = p(w) \prod_{i=1}^{N} p_i(y_i | x_i, w) \tag{7}$$

The goal is to calculate the posterior distribution $p(w|y, x)$. Variational inference algorithm consists in minimizing the Kullback Leibler (KL) divergence between a candidate family of parametric distributions $\{q(w, \theta), \theta \in \Theta \subset \mathbb{R}^d\}$ and the posterior distribution $p(w|y, x)$ of the global latent variable w. In most implementations, $q(w; \theta)$ belongs to a simple family of distributions such as the multivariate Gaussian family with mean ρ and covariance matrix $\sigma^2 \operatorname{Id}$ in which case $\theta = (\rho, \sigma^2) \in \Theta = \mathbb{R} \times \mathbb{R}_+^*$. The variational inference problem boils down to minimizing the following KL divergence:

$$\theta^* = \arg\min_{\theta \in \Theta} \operatorname{KL}(q(w; \theta) \parallel p(w|y, x)) = \arg\min_{\theta \in \Theta} f(\theta) \tag{8}$$

where for all $\theta \in \Theta$, $f(\theta) = \sum_{i=1}^{N} f_i(\theta)$ with :

$$f_i(\theta) \triangleq - \int_{\mathsf{W}} q(w; \theta) \log p_i(y_i, x_i | w) \mathrm{d}w + \frac{1}{N} \operatorname{KL}(q(w; \theta) \parallel p(w)) = r_i(\theta) + d(\theta) \tag{9}$$

Even though this procedure makes inference analytical for a large class of models, it still lacks in many ways. This technique does not scale to large data (evaluating the reconstruction term (9) requires calculations over the entire dataset) and the approach does not adapt to complex models (models in which this last integral cannot be evaluated analytically) such as Bayesian neural networks [Neal, 2012, Gal, 2016]. Monte Carlo integration and mini-batch strategies, as in [Hoffman et al., 2013, Titsias and Lázaro-Gredilla, 2014, Kucukelbir et al., 2017, Kingma and Welling, 2013] are thus preferred here. Optimization of this criterion can be performed using our incremental stochastic surrogate optimization framework. We use the following quadratic surrogate at $\theta \in \Theta$:

$$f_{i,\theta}(\vartheta) \triangleq f_i(\theta) + \nabla f_i(\theta)^\top (\vartheta - \theta) + \frac{L}{2} \|\vartheta - \theta\|_2^2 \tag{10}$$

where $\|\cdot\|_2$ is the ℓ_2-norm and L is an upper bound of the spectral norm of the Hessian of f_i at θ. The gradient $\nabla f_i(\theta)$ can be computed several ways [Paisley et al., 2012]. We use the reparametrization trick suggested in [Kingma and Welling, 2013, Blundell et al., 2015]. For $\theta \in \Theta$ and $e \in \mathbb{R}^d$, let $t : \Theta \times \mathbb{R}^d \mapsto \mathbb{R}^d$ be a function and ϕ be the density of the standard multivariate normal distribution $\mathcal{N}_d(0, \operatorname{Id})$. We assume that for all $\theta \in \Theta$, the distribution of the random vector $W = t(\theta, \epsilon)$ where $\epsilon \sim \mathcal{N}_d(0, \operatorname{Id})$ has a density $q(\cdot, \theta)$. Then, following [Blundell et al., 2015, Proposition 1]:

$$\nabla_\theta \int_{\mathsf{W}} \log p_i(y_i, x_i | w) q(w, \theta) \mathrm{d}w = \int_{\mathsf{W}} \operatorname{J}(\theta, e) \nabla_w \log p_i(y_i, x_i | t(\theta, e)) \phi(e) \mathrm{d}e$$

where for each $e \in \mathbb{R}^d$, $\operatorname{J}(\theta, e)$ is the Jacobian of the function $t(\cdot, e)$ with respect to θ. Consequently, the pair $(r_{i,\theta}(e, \vartheta), \phi(e))$ defining $f_{i,\theta}(\vartheta)$ is given by:

$$r_{i,\theta}(e, \vartheta) \triangleq (-\log p_i(y_i, x_i | t(\theta, e)) + d(\theta))$$
$$+ (-\operatorname{J}(\theta, e) \nabla_\theta \log p_i(y_i, x_i | t(\theta, e)) + \nabla d(\theta))^\top (\vartheta - \theta) + \frac{L}{2} \|\vartheta - \theta\|_2^2 \tag{11}$$

Our scheme is based on the computation, at each iteration, of stochastic auxiliary functions for a mini-batch of components. For $i \in [\![N]\!]$, the auxiliary function, noted $\hat{f}_{i,\theta}(\vartheta)$ is a Monte Carlo approximation of the surrogate function $f_{i,\theta}(\vartheta)$ defined by (5) such that:

$$\hat{f}_{i,\theta}(\vartheta) \triangleq \frac{1}{M} \sum_{m=0}^{M-1} r_{i,\theta}(z_i^m, \vartheta) \quad \text{for all } (\theta, \vartheta) \in \Theta^2 \tag{12}$$

where $\{z_i^m\}_{m=0}^{M-1}$ is a Monte Carlo batch. In simple scenarios, the samples $\{z_i^m\}_{m=0}^{M-1}$ are conditionally independent and identically distributed with distribution $p_i(z_i, \theta)$. Nevertheless, in many cases, sampling exactly from this distribution is not an option and the Monte Carlo batch is sampled by

Monte Carlo Markov Chains (MCMC) algorithm. The MISSO algorithm can be summarized as follows:

Algorithm 2 MISSO algorithm

Initialization: given an initial parameter estimate θ^0, for all $i \in [\![N]\!]$ compute the function $\vartheta \to \hat{f}_{i,\theta^0}(\vartheta)$ defined by (12).

Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![N]\!], \mathrm{card}(A) = p\}$

2. For all $i \in I_k$, sample a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ from $p_i(z_i, \theta^{k-1})$.

3. For all $i \in I_k$, compute the function $\vartheta \to \hat{f}_{i,\theta^{k-1}}(\vartheta)$ defined by (12).

4. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^{N} \hat{a}_i^k(\vartheta)$ where $\hat{a}_i^k(\vartheta)$ are defined recursively as follows:

$$\hat{a}_i^k(\vartheta) \triangleq \begin{cases} \hat{f}_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ \hat{a}_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{13}$$

Whether we use Markov Chain Monte Carlo or direct simulation, we need to control the supremum norm of the fluctuations of the Monte Carlo approximation. Let $i \in [\![N]\!]$, $\{j_i(z_i, \vartheta), z_i \in \mathsf{Z}_i, \vartheta \in \Theta\}$ be a family of measurable functions, λ_i a probability measure on $\mathsf{Z}_i \times \mathcal{Z}_i$. We define:

$$C_i(j_i) \triangleq \sup_{\theta \in \Theta} \sup_{M > 0} M^{-1/2} \mathbb{E}_{i,\theta} \left[\sup_{\vartheta \in \Theta} \left| \sum_{m=0}^{M-1} \left\{ j_i(z_i^m, \vartheta) - \int_{\mathsf{Z}_i} j_i(z_i, \vartheta) p_i(z_i, \theta) \lambda_i(\mathrm{d}z_i) \right\} \right| \right] \tag{14}$$

M 3. *For all $i \in [\![N]\!]$ and $\theta \in \Theta$:*

$$\lim_{k \to \infty} C_i(r_{i,\theta}) < \infty \quad \text{and} \quad \lim_{k \to \infty} C_i(\nabla r_{i,\theta}) < \infty. \tag{15}$$

When this approximation is done using an MCMC procedure to perform a Monte Carlo integration, the assumption M 3 is based on maximal inequality for beta-mixing sequences obtained in [Doukhan et al., 1995]. This condition can be translated in terms of drift and minorization conditions (see [Meyn and Tweedie, 2009]). Finally, we consider the following assumption on the number of simulations:

M 4. *$\{M_k\}_{k \geq 0}$ is a non deacreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$.*

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

$$V_k \leq V_{k-1} - X_k + E_k \tag{16}$$

then:

(i) *for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k \geq 0}$ converges a.s. to a finite limit V_∞.*

(ii) *the sequence $(\mathbb{E}[V_k])_{k \geq 0}$ converges and $\lim_{k \to \infty} \mathbb{E}[V_k] = \mathbb{E}[V_\infty]$.*

(iii) *the series $\sum_{k=0}^{\infty} X_k$ converges almost surely and $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.*

Proof. The proof is postponed to the appendix $\qquad \square$

Remark 1. *Note that the result still holds if $(V_k)_{k \geq 0}$ is a sequence of random variables which is bounded from below by a deterministic quantity $M \in \mathbb{R}$.*

Theorem 1. *Assume M1-M4. Let $(\theta^k)_{k \geq 1}$ be a sequence generated from $\theta^0 \in \Theta$ by the iterative application described by Algorithm 2. Then:*

(i) *$(f(\theta^k))_{k \geq 1}$ converges almost surely.*

(ii) $\left(\theta^k\right)_{k \geq 1}$ *satisfies the Asymptotic Stationary Point Condition.*

Proof. The proof is postponed to the appendix □

Example (Incremental MCEM). *In most cases, the surrogate of the incremental EM algorithm defined as:*

$$f_{i,\theta}(\vartheta) \triangleq \int_{\mathsf{Z}_i} \log \frac{p_i(z_i, \theta)}{c_i(z_i, \vartheta)} p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \quad \text{for all } i \in [\![N]\!] \text{ and } (\theta, \vartheta) \in \Theta^2. \tag{17}$$

is intractable. With our notations, we define the Monte Carlo approximation of this surrogate as:

$$\hat{f}_{i,\theta}(\vartheta) \triangleq \frac{1}{M} \sum_{m=0}^{M-1} \log \frac{p_i(z_i^m, \theta)}{c_i(z_i^m, \vartheta)} \quad \text{for all } i \in [\![N]\!] \text{ and } (\theta, \vartheta) \in \Theta^2. \tag{18}$$

where $\{z_i^m\}_{m=0}^{M-1}$ *is a Monte Carlo batch sampled from* $p_i(z_i, \theta)$ *using an MCMC procedure. The MISSO algorithm coincides with the mini-batch version of the MCEM algorithm which yields, at iteration* k, *the following update of the parameter:*

$$\theta^k \in \arg\min_{\vartheta \in \Theta} -\sum_{i=0}^{N} \frac{1}{M_{\tau_{i,k}}} \sum_{m=0}^{M_{\tau_{i,k}}-1} \log c_i(z_i^{\tau_{i,k}+1,m}, \vartheta) \tag{19}$$

where $\{z_i^{\tau_{i,k}+1,m}\}_{m=0}^{M-1}$ *is a Monte Carlo batch sampled from* $p_i(z_i, \theta^{\tau_{i,k}})$.

Example (Incremental Variational Inference for latent data model). *The MISO surrogate defined for all* $(\vartheta, \theta) \in \Theta^2$ *by the pair* $(r_{i,\theta}(e, \vartheta), \phi(e))$ *with* $r_{i,\theta}(e, \vartheta)$ *defined by* (11) *is often intractable. The MISSO algorithm coincides with a mini-batch version of the Variational Inference algorithm. At iteration* k, *and given the quadratic structure of the surrogate (see* (11)*), the MISSO algorithm consists in:*

1. *Picking a set* I_k *uniformly on* $\{A \subset [\![N]\!], \operatorname{card}(A) = p\}$.

2. *Sampling a Monte Carlo batch* $\{\epsilon^{k,m}\}_{m=0}^{M_k-1}$ *from the standard Gaussian distribution.*

3. *Setting* $\theta^k = \theta^{k-1} - \frac{1}{L}\sum_{i=1}^{N} \hat{a}_i^k$ *where* $\hat{a}_i^k$ *are defined recursively as follows:*

$$\hat{a}_i^k \triangleq \begin{cases} -\frac{1}{M_k}\sum_{m=0}^{M_k-1} \mathrm{J}(\theta, e^{k,m})\nabla_\theta \log p_i(y_i, x_i | t(\theta, e^{k,m})) + \nabla d(\theta^{k-1}) & \text{if } i \in I_k \\ \hat{a}_i^{k-1} & \text{otherwise} \end{cases} \tag{20}$$

3 Numerical Applications

3.1 Fitting a logistic regression for a binary variable

The model. Let $y = (y_i, i \in [\![N]\!])$ be the vector of binary responses where for each individual i, $y_i = (y_{ij}, 1 \leq j \leq n_i)$ is a sequence of conditionally independent random variables taking values in $\{0, 1\}$ which corresponds to the j-th responses for the i-th subject. We consider a logistic regression problem in which the parameters depend upon each individual i. Denote by $z_i = (z_{i,p}) \in \mathbb{R}^p$ the vector of regression coefficients (the latent data) for individual i and $((d_{ij}), 1 \leq j \leq n_i)$ the associated explanatory variables. The conditional distribution of the observations y_i given the latent variables z_i is given by:

$$\operatorname{logit}(\mathbb{P}(y_{ij} = 0 | z_i)) = d_{ij}^\top z_i$$

For all $i \in [\![N]\!]$, we assume that z_i are independedn and marginally distributed according to $\mathcal{N}(\beta, \Omega)$. The complete log-likelihood is expressed as:

$$\log c(z, \theta) \propto \sum_{i=1}^{N} \sum_{j=1}^{n_i} \{y_{ij} d_{ij}^\top z_i - \log(1 + e^{d_{ij}^\top z_i})\} - \sum_{i=1}^{N} \frac{1}{2}\log(|\Omega|) + \frac{1}{2}\operatorname{Tr}\left(\Omega^{-1}(z_i - \beta)(z_i - \beta)^\top\right)$$

We want to compute the maximum likelihood estimator for the parameter θ which maximizes the incomplete likelihood $\int_Z c(z, \theta) \prod_{i=1}^N \phi(z_i; \beta, \Omega) \mathrm{d}z_i$ where $\phi(z_i; \beta, \Omega)$ is the density of a multivariate Gaussian variable with mean β and covariance Ω. Since the expectation of the complete log likelihood with respect to the conditional distribution of the latent variables given the observations is intractable, we use the MISSO algorithm. Computing the surrogates, defined by (18), requires to simulate random draws from this conditional distribution. For this purpose, we use the saemix R package [Comets et al., 2017] to run a Metropolis-Hastings within Gibbs sampler [Brooks et al., 2011]. Furthermore, this model belongs to the curved exponential family [Keener, 2010] where for all $i \in [\![N]\!]$; the complete data sufficient statistics are given by $\tilde{S}_i(z_i) \triangleq (z_i, z_i^\top z_i)$. At the k-th iteration, the MISSO algorithm consists in picking a set I_k, sampling a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ for $i \in I_k$ and computing the quantities $(s_i^{1,k}, s_i^{2,k})$ as follows:

$$(s_i^{1,k}, s_i^{2,k}) = \begin{cases} \left(\frac{1}{M_k} \sum_{m=0}^{M_k-1} z_i^{k,m}, \frac{1}{M_k} \sum_{m=0}^{M_k-1} (z_i^{k,m})^\top z_i^{k,m} \right) & \text{if } i \in I_k \\ (s_i^{1,k-1}, s_i^{2,k-1}) & \text{otherwise} \end{cases} \quad (21)$$

and finally setting $\beta^k = \frac{1}{N} \sum_{i=1}^N s_i^{1,k}$ and $\Omega^k = \frac{1}{N} \sum_{i=1}^N s_i^{2,k} - (\beta^k)^\top \beta^k$ (see section 2 of the appendix material for more details).

Simulation and runs. In the sequel, $p = 3$, $N = 1200$ and for all $i \in [\![N]\!]$, $n_i = 15$. For all $i \in [\![N]\!]$ and $j \in [\![n_i]\!]$, we take $d_{ij,1} = 1$, $d_{ij,2} = -20 + (j-1) * 5$ and for $i \in [\![N]\!]$ $d_{ij,3} = 10\lceil 3i/N \rceil$. The data are generated using the following values for the fixed and random effects $(\beta_1 = -4, \beta_2 = -0.5, \beta_3 = 1, \omega_1 = 0.3, \omega_2 = 0.2, \omega_3 = 0.2)$. The size of the Monte Carlo batch increases polynomially, $M_k \triangleq M_0 + k^2$ with $M_0 = 50$. Figure 1 shows the convergence of the fixed effects $(\beta_1, \beta_2, \beta_3)$ estimates obtained with both the MCEM and the mini-batch MCEM algorithms using our MISSO scheme (19) for different batch sizes.
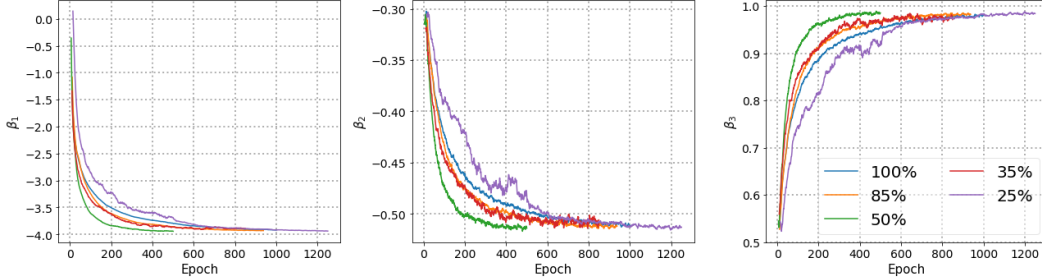

Figure 1: (Incremental MCEM) Convergence of the vector of fixed parameters β for different batch sizes function of passes over the data.

3.2 Variational inference for Bayesian neural network

In this section, we apply variational inference for a 2-layer Bayesian neural network on the MNIST dataset [LeCun and Cortes, 2010] with our MISSO scheme. The training set is composed of $N = 60\,000$ handwritten digits, 28×28 images, $d = 784$. Our neural network is composed of an input layer with $d = 784$ units, a single hidden layer of $p = 100$ hyperbolic tangent units and a final softmax output layer with $K = 10$ classes.

We use the framework developed in Example 2.1 with $p(w) = \mathcal{N}(0, \mathrm{Id})$ and $p(y_i|x_i, w) = \text{Softmax}(f(x_i, w))$ where f is the two layer model described above. The variational distribution $q(w, \theta)$ is set to be the multivariate Gaussian distribution $\mathcal{N}(\rho, \sigma^2 \mathrm{Id})$. At the k-th iteration, the update of the MISSO algorithm is given by (20).

We compare the convergence behaviors of the following state of the art optimization algorithms, using their vanilla implementations on TensorFlow [Abadi et al., 2015]: the SGD [Kiefer and Wolfowitz, 1952], the ADAM [Kingma and Ba, 2014] and the Momentum [Sutskever et al., 2013] algorithms versus our MISSO update with a constant learning rate of 10^{-5}. The loss function (9) and its gradients were computed by Monte Carlo integration using Edward library [Tran et al., 2016], based on the

7

reparametrization trick. We run those algorithms using 1% and 10% of the training set. Figure 2 shows the convergence of the objective function through the epochs. For both mini-batch sizes, our framework does better than SGD and ADAM. Similar rates are observed between MISSO and Momentum which makes sense given the similarities in the update step.
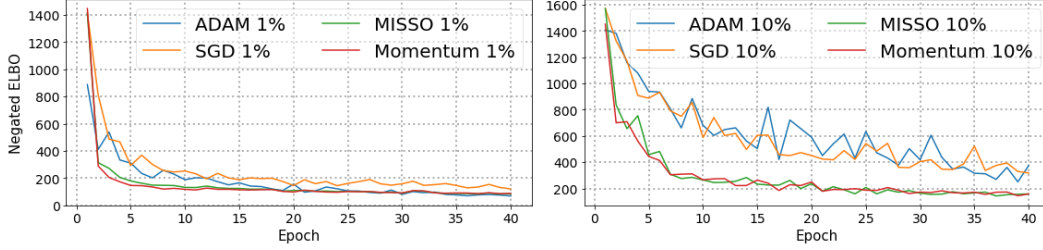

Figure 2: (Incremental Variational Inference) Convergence of the negated ELBO for 40 epochs over the training set. Runs for two different mini-batch sizes (1% left and 10% right).

4 Conclusion

In this paper, we have presented a unifying framework for minimization by incremental surrogate optimization when the surrogate functions are intractable and need to be approximated by Monte Carlo. Our approach covers a large class of nonconvex optimization algorithms used in machine learning, such as mini-batch version of the MCEM and the Variational Bayes inference algorithms. We provided proofs of convergence. Compared to the state-of-the-art algorithms, the incremental approach reduces significantly the variance.

Non asymptotic convergence results for both convex and nonconvex objective functions can be obtained and will be reported in future works.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M Schuster, J Shlens, B Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

R. Turner B. Trippe. Overpruning in Variational Bayesian Neural Networks. 2018.

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

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015. URL `http://dl.acm.org/citation.cfm?id=3045118.3045290`.

L. Bottou, F.E. Curtis, and J. Nocedal. Optimization methods for large-scale machine learning, 2016. URL `http://arxiv.org/abs/1606.04838`. cite arxiv:1606.04838.

S. Brooks, A. Gelman, G.L. Jones, and X. Meng, editors. *Handbook of Markov chain Monte Carlo*. Chapman & Hall/CRC Handbooks of Modern Statistical Methods. CRC Press, Boca Raton, FL, 2011. ISBN 978-1-4200-7941-8. URL `https://doi.org/10.1201/b10905`.

E. Comets, A. Lavenu, and M. Lavielle. Parameter estimation in nonlinear mixed effect models using saemix, an r implementation of the saem algorithm. *Journal of Statistical Software*, 2017. doi: 10.18637/jss.v080.i03. URL `http://www.hal.inserm.fr/inserm-01502767`. Conditionally accepted for publication on 2017-07-13Estimated delay in publication - one year (editor's communication, 2018-04-05).

S. Mohamed D. J. Rezende and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. 2014.

N. de Freitas, P. Højen-Sørensen, M.I Jordan, and S. Russell. Variational mcmc. In *Proceedings of the Seventeenth Conference on Uncertainty in Artificial Intelligence*, UAI'01, pages 120–127, San Francisco, CA, USA, 2001. Morgan Kaufmann Publishers Inc. ISBN 1-55860-800-1. URL `http://dl.acm.org/citation.cfm?id=2074022.2074038`.

A. Defazio, F. Bach, and S. Lacoste-Julien. SAGA: A fast incremental gradient method with support for non-strongly convex composite objectives. In *NIPS*, pages 1646–1654, 2014.

P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. *Ann. Inst. H. Poincaré Probab. Statist.*, 31(2):393–427, 1995. ISSN 0246-0203. URL `http://www.numdam.org/item?id=AIHPB_1995__31_2_393_0`.

G. Fort and E. Moulines. Convergence of the monte carlo expectation maximization for curved exponential families. *Ann. Statist.*, 31(4):1220–1259, 08 2003. doi: 10.1214/aos/1059655912. URL `https://doi.org/10.1214/aos/1059655912`.

Y. Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521, 2015.

M. Girolami and B. Calderhead. Riemann manifold Langevin and Hamiltonian Monte Carlo methods. *J. R. Stat. Soc. Ser. B Stat. Methodol.*, 73(2):123–214, 2011. ISSN 1369-7412. doi: 10.1111/j.1467-9868.2010.00765.x. URL `https://doi.org/10.1111/j.1467-9868.2010.00765.x`. With discussion and a reply by the authors.

Blei D.M. Hoffman, M. D., C. Wang, and J.W. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1):1303–1347, 2013.

Y. Bengio I. Goodfellow and A. Courville. Deep learning book. *MIT Press*, 2016.

R.W. Keener. *Curved Exponential Families*, pages 85–99. Springer New York, New York, NY, 2010. ISBN 978-0-387-93839-4. doi: 10.1007/978-0-387-93839-4_5. URL `https://doi.org/10.1007/978-0-387-93839-4_5`.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466, 09 1952. doi: 10.1214/aoms/1177729392. URL `https://doi.org/10.1214/aoms/1177729392`.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14`.

D.P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL `http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13`.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18(1):430–474, January 2017. ISSN 1532-4435. URL `http://dl.acm.org/citation.cfm?id=3122009.3122023`.

K. Lange. MM Optimization Algorithms. 2016.

N. Le Roux, M.W. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2672–2680, 2012.

Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL `http://yann.lecun.com/exdb/mnist/`.

J. Mairal. Incremental Majorization-Minimization Optimization with Application to Large-Scale Machine Learning. *SIAM*, 2015.

G. McLachlan and T. Krishnan. The EM Algorithm and Extensions. 2007.

S. Meyn and R.L. Tweedie. *Markov chains and stochastic stability*. Cambridge University Press, Cambridge, second edition, 2009. ISBN 978-0-521-73182-9. URL `https://doi.org/10.1017/CBO9780511626630`. With a prologue by Peter W. Glynn.

V. Mullachery, A. Khera, and A. Husain. Bayesian neural networks. *CoRR*, abs/1801.07710, 2018. URL `http://arxiv.org/abs/1801.07710`.

R. Neal and G.E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

R. M. Neal. Bayesian learning for neural networks. *Springer Science Business Media*, 118, 2012.

R. Neath. On Convergence Properties of the Monte Carlo EM Algorithm. 2012.

Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. URL `https://EconPapers.repec.org/RePEc:cor:louvco:2007076`.

J.W. Paisley, D.M. Blei, and M.I. Jordan. Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress, 2012. URL `http://dblp.uni-trier.de/db/conf/icml/icml2012.html#PaisleyBJ12`.

N. Parikh and S.P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3):127–239, 2014.

N. Pawlowski, M. Jaques, and B. Glocker. Efficient variational Bayesian neural network ensembles for outlier detection. 2017.

N.G. Polson and V.O. Sokolov. Deep Learning: A Bayesian Perspective. 2017.

R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. *PMLR*, 33:814–822, 2014.

M. Razaviyayn, M. Sanjabi, and Z. Luo. A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *CoRR*, abs/1307.4457, 2013.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

I. Sutskever, J. Martens, G. E. Dahl, and G.E. Hinton. On the importance of initialization and momentum in deep learning. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147. JMLR.org, 2013. URL `http://dblp.uni-trier.de/db/conf/icml/icml2013.html#SutskeverMDH13`.

M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1971–1979, Bejing, China, 22–24 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v32/titsias14.html`.

D. Tran, A. Kucukelbir, A.B. Dieng, M. Rudolph, D. Liang, and D.M. Blei. Edward: A library for probabilistic modeling, inference, and criticism, 2016. URL `http://arxiv.org/abs/1610.09787`. cite arxiv:1610.09787.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1:pp. 1–305, 2008.

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.*, 1990.

Y. Gal Y. Li. Dropout Inference in Bayesian Neural Networks with Alpha-divergences. 2017.

A Proofs

A.1 Proof of Lemma 1

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

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \tag{22}$$

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

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

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \tag{23}$$

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

$$W_k \leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \tag{24}$$
$$\mathbb{E}[W_k] \leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k]$$

Hence the sequences $(W_k)_{k \geq 0}$ and $(\mathbb{E}[W_k])_{k \geq 0}$ are non increasing. Since for all $k \geq 0$, $W_k \geq -\sum_{j=1}^{\infty} |E_j| > -\infty$ and $\mathbb{E}[W_k] \geq -\sum_{j=1}^{\infty} \mathbb{E}[|E_j|] > -\infty$, the (random) sequence $(W_k)_{k \geq 0}$ converges a.s. to a limit W_∞ and the (deterministic) sequence $(\mathbb{E}[W_k])_{k \geq 0}$ converges to a limit w_∞. Since $|W_k| \leq V_0 + \sum_{j=1}^{\infty} |E_j|$, the Fatou lemma implies that:

$$\mathbb{E}[\liminf_{k \to \infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k \to \infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \tag{25}$$

showing that the random variable W_∞ is integrable.

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

$$\lim_{k \to \infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k \to \infty} U_k] \tag{26}$$

Finally, we have:

$$\lim_{k \to \infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k \to \infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \tag{27}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (24) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\sum_{j=1}^{\infty} X_j \leq W_0 - W_\infty < \infty$$
$$\sum_{j=1}^{\infty} \mathbb{E}[X_j] \leq \mathbb{E}[W_0] - w_\infty < \infty \tag{28}$$

which concludes the proof of the lemma.

A.2 Proof of theorem 1

A.2.1 Proof of (i)

Set for all $\vartheta \in \Theta$, $i \in [\![N]\!]$ and $k \geq 1$:

$$a_i^k(\vartheta) \triangleq f_{i,\theta^{\tau_{i,k}}}(\vartheta) \quad \text{and} \quad \bar{a}^k(\vartheta) = \sum_{i=1}^N a_i^k(\vartheta) \tag{29}$$

where the function $f_{i,\theta^{\tau_{i,k}}}$ is defined by (5) and $\tau_{i,k}$ is defined by (4). For any $k \geq 1$ and $\theta \in \Theta$ the following decomposition plays a key role:

$$\hat{a}^k(\vartheta) = \hat{a}^{k-1}(\vartheta) + \sum_{i \in I_k} \{\hat{f}_{i,\theta^{k-1}}(\vartheta) - \hat{a}_i^{k-1}(\vartheta)\} \tag{30}$$

where for all $\vartheta \in \Theta$, $i \in [\![N]\!]$ and $k \geq 1$:

$$\hat{a}_i^k(\vartheta) \triangleq \hat{f}_{i,\theta^{\tau_{i,k}}}(\vartheta) \quad \text{and} \quad \hat{a}^k(\vartheta) = \sum_{i=1}^N \hat{a}_i^k(\vartheta) \tag{31}$$

Set the following notations:

$$V_k \triangleq \bar{a}^k(\theta^k),$$

$$X_k \triangleq - \sum_{i \in I_k} \{f_{i,\theta^{k-1}}(\theta^{k-1}) - a_i^{k-1}(\theta^{k-1})\},$$

$$E_k \triangleq \sum_{i \in I_k} \{\hat{f}_{i,\theta^{k-1}}(\theta^{k-1}) - f_{i,\theta^{k-1}}(\theta^{k-1})\}$$

$$+ \sum_{i \in I_k} \{a_i^{k-1}(\theta^{k-1}) - \hat{a}_i^{k-1}(\theta^{k-1})\}$$

$$+ \bar{a}^k(\theta^k) - \hat{a}^k(\theta^k) + \hat{a}^{k-1}(\theta^{k-1}) - \bar{a}^{k-1}(\theta^{k-1}).$$

Combining (30) with $\bar{a}^k(\theta^k) = \bar{a}^k(\theta^k) - \hat{a}^k(\theta^k) + \hat{a}^k(\theta^k)$ and $\hat{a}^k(\theta^k) \leq \hat{a}^k(\theta^{k-1})$, we obtain:

$$V_k \leq V_{k-1} - X_k + E_k. \tag{32}$$

where a_i^{k-1} and $\bar{a}^k$ are defined in (29). We now check the assumptions of Lemma 1. Note first that the sequence $(V_k)_{k\geq0}$ is bounded from below under assumption M 2. We now check that $X_k \geq 0$ thanks to the following relation obtained using the definition of surrogate functions:

$$X_k = \sum_{i \in I_k} \{a_i^{k-1}(\theta^{k-1}) - f_{i,\theta^{k-1}}(\theta^{k-1})\} = \sum_{i \in I_k} \{a_i^{k-1}(\theta^{k-1}) - f_i(\theta^{k-1})\} \geq 0. \tag{33}$$

We finally have to prove the convergence of the series $\sum_{k=0}^\infty \mathbb{E}[|E_k|]$. For this purpose, we will show that for all $i \in [\![N]\!]$:

$$\sum_{k=0}^\infty \mathbb{E}\big[|\hat{a}_i^k(\theta^k) - a_i^k(\theta^k)|\big] < \infty \tag{34}$$

We have, using the Tower property of the conditional expectation and the Jensen inequality:

$$\mathbb{E}\big[|\hat{a}_i^k(\theta^k) - a_i^k(\theta^k)|\big] \leq \mathbb{E}\left[\mathbb{E}_{i,\theta^{\tau_{i,k}}}\left[\sup_{\vartheta \in \Theta} |\hat{a}_i^k(\vartheta) - a_i^k(\vartheta)|\right]\right] \tag{35}$$

Under assumption M 3 applied with the function $\vartheta \to \hat{a}_i^k(\vartheta)$, for all $i \in [\![N]\!]$ we have:

$$\mathbb{E}_{i,\theta^{\tau_{i,k}}}\left[\sup_{\vartheta \in \Theta} |\hat{a}_i^k(\vartheta) - a_i^k(\vartheta)|\right] \leq C_i(r_{i,\theta^{\tau_{i,k}}}) M_{\tau_{i,k}}^{-1/2} \tag{36}$$

where $C_i(r_{i,\theta^{\tau_{i,k}}})$ is a finite constant defined by (14) and $\tau_{i,k}$ is defined by (4). Thus, we have that:

$$\mathbb{E}\big[|\hat{a}_i^k(\theta^k) - a_i^k(\theta^k)|\big] \leq C_i(r_{i,\theta^{\tau_{i,k}}}) \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] \tag{37}$$

13

Since, any index i is included in a mini-batch with a probability equal to $\frac{p}{N}$ conditionally independently from the past, we obtain that:

$$\mathbb{E}[M_{\tau_{i,k}}^{-1/2}] = \sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \tag{38}$$

Taking the infinite sum of this term yields:

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] &= \sum_{k=1}^{\infty} \sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \\
&= \sum_{k=1}^{\infty} \sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{k-(l+1)} \frac{p}{N} \mathbb{1}_{\{l \leq k-1\}} M_l^{-1/2} \\
&= \frac{p}{N} \sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{-(l+1)} M_l^{-1/2} \sum_{k=l+1}^{\infty} \left(1 - \frac{p}{N}\right)^{k} \\
&= \sum_{l=0}^{\infty} M_l^{-1/2}
\end{aligned} \tag{39}$$

which proves identity (34), using assumption M 4. By summing over the indices $i \in [\![N]\!]$, (34) implies:

$$\sum_{k=0}^{\infty} \mathbb{E}\left[|\hat{a}^k(\theta^k) - \bar{a}^k(\theta^k)|\right] < \infty \tag{40}$$

Hence, we obtain that $\sum_{k=0}^{\infty} |\hat{a}^k(\theta^k) - \bar{a}^k(\theta^k)| < \infty$ almost surely which implies that:

$$\lim_{k \to \infty} \hat{a}^k(\theta^k) - \bar{a}^k(\theta^k) = 0 \quad \text{a.s.} \tag{41}$$

Similarly, using assumption M 3 applied for all $i \in [\![N]\!]$, with the function $\vartheta \to \nabla \hat{a}_i^k(\vartheta)$ we obtain:

$$\lim_{k \to \infty} \nabla \hat{a}^k(\theta^k) - \nabla \bar{a}^k(\theta^k) = 0 \quad \text{a.s.} \tag{42}$$

It follows from (34) and (40) that $\sum_{k=0}^{\infty} \mathbb{E}\left[|E_k|\right] < \infty$ and that the series $\sum_{k=0}^{\infty} \epsilon_k$ converges to an almost surely finite limit. Hence by Lemma 1 and (41) we get:

- the sequence $\left(\bar{a}^k(\theta^k)\right)_{k \geq 0}$ and the series $\sum_{k=0}^{\infty} \chi_k$ converge a.s.
- the sequence $\left(\mathbb{E}\left[\bar{a}^k(\theta^k)\right]\right)_{k \geq 0}$ and the series $\sum_{k=0}^{\infty} \mathbb{E}\left[X_k\right]$ converge with $\lim_{k \to \infty} \mathbb{E}\left[\bar{a}^k(\theta^k)\right] = \mathbb{E}[\lim_{k \to \infty} \bar{a}^k(\theta^k)]$.
- the sequence $\left(\hat{a}^k(\theta^k)\right)_{k \geq 0}$ converges a.s. and the sequence $\left(\mathbb{E}\left[\hat{a}^k(\theta^k)\right]\right)_{k \geq 0}$ converges.

Now, we have to prove the almost-sure convergence of the sequence $\left(f(\theta^k)\right)_{k \geq 0}$ and the convergence of $\left(\mathbb{E}\left[f(\theta^k)\right]\right)_{k \geq 0}$.

Let us denote for all $\theta \in \Theta$ and a subset $J \subset [\![N]\!]$:

$$\begin{aligned}
f_J(\theta) &\triangleq \sum_{i \in J} f_i(\theta) \\
a_J^{k-1}(\theta) &\triangleq \sum_{i \in J} a_i^{k-1}(\theta)
\end{aligned} \tag{43}$$

The Beppo-Levi theorem and the Tower property of the conditional expectation imply:

$$\begin{aligned}
\mathsf{M} \triangleq \mathbb{E}\left[\sum_{k=1}^{\infty} X_k\right] &= \sum_{k=0}^{\infty} \mathbb{E}\left[a_{I_k}^{k-1}(\theta^{k-1}) - f_{I_k}(\theta^{k-1})\right] \\
&= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}\left[a_{I_k}^{k-1}(\theta^{k-1}) - f_{I_k}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right]\right]
\end{aligned} \tag{44}$$

14

with $\mathbb{E}\left[f_{I_k}(\theta^{k-1})\,\middle|\,\mathcal{F}_{k-1}\right] = \frac{p}{N}f(\theta^{k-1})$ and $\mathbb{E}\left[[a_{I_k}^{k-1}(\theta^{k-1})\,\middle|\,\mathcal{F}_{k-1}\right] = \frac{p}{N}\sum_{i=1}^{N}a_i^{k-1}(\theta^{k-1}) = \frac{p}{N}\bar{a}^{k-1}(\theta^{k-1})$ where $\mathcal{F}_{k-1} = \sigma(I_j, j \leq k-1)$ is the filtration generated by the sampling of the indices. We thus obtain:

$$\mathsf{M} = \frac{p}{N}\sum_{k=0}^{\infty}\mathbb{E}\left[\bar{a}^{k-1}(\theta^{k-1}) - f(\theta^{k-1})\right] = \frac{p}{N}\mathbb{E}\left[\sum_{k=0}^{\infty}\bar{a}^{k-1}(\theta^{k-1}) - f(\theta^{k-1})\right] < \infty \quad (45)$$

which yields to:

$$\mathbb{E}\left[\sum_{k=1}^{\infty}X_k\right] = \frac{p}{N}\mathbb{E}\left[\sum_{k=1}^{\infty}\{\bar{a}^{k-1}(\theta^{k-1}) - f(\theta^{k-1})\}\right] < \infty \quad (46)$$

showing that:

$$\begin{aligned}
&\lim_{k\to\infty}\mathbb{E}\left[\bar{a}^k(\theta^k) - f(\theta^k)\right] = 0 \\
&\lim_{k\to\infty}\bar{a}^k(\theta^k) - f(\theta^k) = 0 \quad \text{a.s.}
\end{aligned} \quad (47)$$

showing that the sequence $\left(\mathbb{E}\left[f(\theta^k)\right]\right)_{k\geq 0}$ converges and that $\left(f(\theta^k)\right)_{k\geq 0}$ converges a.s.

A.2.2 Proof of (ii)

Let us define, for all $k \geq 0$, $\bar{h}_k$ as:

$$\bar{h}^k : \vartheta \to \sum_{i=1}^{N}a_i^k(\vartheta) - f_i(\vartheta) \quad (48)$$

$\bar{h}^k$ is L-smooth with $L = \sum_{i=1}^{N}L_i$ since each of its component is L_i-smooth by definition of the surrogate functions. Using the particular parameter $\vartheta^k = \theta^k - \frac{1}{L}\nabla\bar{h}_k(\theta^k)$ we have the following classical inequality for smooth functions (cf. Lemma 1.2.3 in [Nesterov, 2007]):

$$\begin{aligned}
&0 \leq \bar{h}^k(\vartheta^k) \leq \bar{h}^k(\theta^k) - \frac{1}{2L}\|\nabla\bar{h}^k(\theta^k)\|_2^2 \\
&\implies \|\nabla\bar{h}^k(\theta^k)\|_2^2 \leq 2L\bar{h}^k(\theta^k)
\end{aligned} \quad (49)$$

Using (47), we conclude that $\lim_{k\to\infty}\|\nabla\bar{h}^k(\theta^k)\|_2 = 0$ a.s. Then, the decomposition of $\langle\nabla f(\theta^k), \theta - \theta^k\rangle$ for any $\theta \in \Theta$ yields:

$$\begin{aligned}
\langle\nabla f(\theta^k), \theta - \theta^k\rangle &= \langle\nabla\bar{a}^k(\theta^k), \theta - \theta^k\rangle - \langle\nabla\bar{h}^k(\theta^k), \theta - \theta^k\rangle \\
&= \langle\nabla\bar{a}^k(\theta^k) - \nabla\hat{a}^k(\theta^k), \theta - \theta^k\rangle + \langle\nabla\hat{a}^k(\theta^k), \theta - \theta^k\rangle - \langle\nabla\bar{h}^k(\theta^k), \theta - \theta^k\rangle
\end{aligned} \quad (50)$$

Note that θ^k is the result of the minimization of $\hat{a}^k(\theta)$ on the constrained set Θ, therefore for all $\theta \in \Theta$, $\langle\nabla\hat{a}^k(\theta^k), \theta - \theta^k\rangle \geq 0$. Thus, we obtain, using the Cauchy-Schwarz inequality:

$$\begin{aligned}
\langle\nabla f(\theta^k), \theta - \theta^k\rangle &\geq \langle\nabla\bar{a}^k(\theta^k) - \nabla\hat{a}^k(\theta^k), \theta - \theta^k\rangle - \langle\nabla\bar{h}^k(\theta^k), \theta - \theta^k\rangle \\
&\geq -\|\nabla\bar{a}^k(\theta^k) - \nabla\hat{a}^k(\theta^k)\|_2\|\theta - \theta^k\|_2 - \|\nabla\bar{h}^k(\theta^k)\|_2\|\theta - \theta^k\|_2
\end{aligned} \quad (51)$$

By minimizing over Θ and taking the infimum limit, we get, using (42):

$$\liminf_{k\to\infty}\inf_{\theta\in\Theta}\frac{\langle\nabla f(\theta^k), \theta - \theta^k\rangle}{\|\theta - \theta^k\|_2} \geq -\lim_{k\to\infty}\left(\|\nabla\bar{a}^k(\theta^k) - \nabla\hat{a}^k(\theta^k)\|_2 + \|\nabla\bar{h}^k(\theta^k)\|_2\right) = 0 \quad (52)$$

which is the Asymptotic Stationary Point Condition (ASPC).

B   Incremental MCEM for Curved Exponential Family

In the particular case where for all $i \in [\![N]\!]$ and $z_i \in \mathsf{Z}_i$, the complete model $\theta \to c_i(z_i, \theta)$ belongs to the curved exponential family, we assume that:

E 1. *For all $i \in [\![N]\!]$ and $\theta \in \Theta$:*

$$\log c_i(z_i, \theta) = H_i(z_i) - \psi_i(\theta) + \langle \tilde{S}_i(z_i), \phi_i(\theta) \rangle. \tag{53}$$

where $\psi_i : \Theta \mapsto \mathbb{R}$ and $\phi_i : \Theta \mapsto \mathbb{R}$ are twice continuously differentiable functions of θ, $H_i : \mathsf{Z}_i \mapsto \mathbb{R}$ is a twice continuously differentiable function of z_i and $\tilde{S}_i : \mathsf{Z}_i \mapsto \mathsf{S}_i$ is a statistic taking its values in a convex subset S_i of $\mathbb{R}$ and such that $\int_{\mathsf{Z}_i} |\tilde{S}_i(z_i)| p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) < \infty$.

Define, for all $\theta \in \Theta$ and $s = (s_i, 1 \le i \le N) \in \mathsf{S}$ where $\mathsf{S} = \bigtimes_{n=1}^{N} \mathsf{S}_i$, the function $L(s; \theta)$ by:

$$L(s; \theta) \triangleq \sum_{i=1}^{N} \psi_i(\theta) - \sum_{i=1}^{N} \langle s_i, \phi_i(\theta) \rangle. \tag{54}$$

E 2. *There exist a function $\hat{\theta} : \mathsf{S} \mapsto \Theta$ such that for all $s \in \mathsf{S}$, :*

$$L(s; \hat{\theta}(s)) \le L(s; \theta). \tag{55}$$

In many models of practical interest for all $s \in \mathsf{S}$, $\theta \mapsto L(s, \theta)$ has a unique minimum. In the context of the curved exponential family, the MISSO algorithm can be formulated as follows:

Algorithm 3 MISSO for a curved exponential family

Initialization: given an initial parameter estimate θ^0, for all $i \in [\![N]\!]$ and $m \in [\![0, M_0 - 1]\!]$, sample a Monte Carlo batch $\{z_i^{0,m}\}_{m=0}^{M_0-1}$ from $p_i(z_i, \theta^0)$ and compute $s_i^0 = \frac{1}{M_0} \sum_{m=1}^{M_0} \tilde{S}_i(z_i^{0,m})$.

Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![N]\!], \mathrm{card}(A) = p\}$

2. For all $i \in I_k$ and $m \in [\![0, M_k - 1]\!]$, sample a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ from $p_i(z_i, \theta^{k-1})$.

3. Compute s_i^k such as:

$$s_i^k = \begin{cases} \frac{1}{M_k} \sum_{m=1}^{M_k - 1} \tilde{S}_i(z_i^{k,m}) & \text{if } i \in I_k \\ s_i^{k-1} & \text{otherwise} \end{cases} \tag{56}$$

4. Set $\theta^k = \hat{\theta}(s^k)$ where $s^k = (s_i^k, 1 \le i \le N)$

In the context of the logisitc regression described in section 3.1, the complete log likelihood is expressed as:

$$\log c(z, \theta) \propto \sum_{i=1}^{N} \sum_{j=1}^{n_i} \{y_{ij} d_{ij}^\top z_i - \log(1 + e^{d_{ij}^\top z_i})\} - \sum_{i=1}^{N} \frac{1}{2} \log(|\Omega|) + \frac{1}{2} \mathrm{Tr}\left(\Omega^{-1}(z_i - \beta)(z_i - \beta)^\top\right)$$

and for all $i \in [\![N]\!]$, the sufficient statistics are defined as $\tilde{S}_i(z_i) \triangleq (z_i, z_i^\top z_i)$. Then, it can easily be shown that the maximization function is defined as follows:

$$\hat{\theta} : \mathsf{S} \mapsto \Theta \tag{57}$$

$$(s_{i,1}, s_{i,2})_{i=1}^{N} \to \left(\frac{1}{N} \sum_{i=1}^{N} s_{i,1}, \frac{1}{N} \sum_{i=1}^{N} s_{i,2} - (\frac{1}{N} \sum_{i=1}^{N} s_{i,1})^\top \frac{1}{N} \sum_{i=1}^{N} s_{i,1} \right) \tag{58}$$