A Doubly Stochastic Surrogate Optimization Scheme for Non-convex Finite-sum Problems

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

Hoi-To Wai
Department of SEEM
The Chinese University of Hong Kong
Hong Kong
htwai@se.cuhk.edu.hk

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

Abstract

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

1 Introduction

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

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

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

To tackle the optimization problem (1), a popular approach is to apply the majorization-minimization (MM) method which iteratively minimizes a majorizing surrogate function. A large number of existing procedures fall into this general framework, for instance gradient-based or proximal methods or the Expectation-Maximization (EM) algorithm [McLachlan and Krishnan, 2008] and some variational Bayes inference techniques [Jordan et al., 1999]; see for example [Razaviyayn et al., 2013] and [Lange, 2016] and the references therein. When the number of terms n in (1) is large, the vanilla MM method may be intractable because it requires to construct a surrogate function for all the n terms $\mathcal{L}_i$ at each iteration. Here, a remedy is to apply the Minimization by Incremental Surrogate Optimization (MISO) method proposed by Mairal [2015], where the surrogate functions are

Preprint. Under review.

updated incrementally. The MISO method can be interpreted as a combination of MM and ideas which have emerged for variance reduction in stochastic gradient methods [Schmidt et al., 2017].

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

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

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

In Section 2, we review the techniques for incremental minimization of finite sum functions based on the MM principle; specifically, we review the MISO method as introduced in [Mairal, 2015], and present a class of surrogate functions expressed as an expectation over a latent space. The MISSO method is then introduced for the latter class of surrogate functions. In Section 3, we provide the asymptotic and non-asymptotic convergence analysis for the MISSO method. Finally, Section 4 presents numerical applications to illustrate our findings including parameter inference for logistic regression with missing covariates and variational inference for Bayesian neural network.

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

$$f'(\boldsymbol{\theta}, \boldsymbol{d}) := \lim_{t \to 0^+} \frac{f(\boldsymbol{\theta} + t\boldsymbol{d}) - f(\boldsymbol{\theta})}{t} \,. \tag{2}$$

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

2 Incremental Minimization of Finite Sum Non-convex Functions

The objective function in (1) is composed of a finite sum of possibly non-smooth and non-convex functions. A popular approach here is to apply the MM method. The MM method tackles (1) through alternating between two steps — (i) minimizing a *surrogate* function which upper bounds the original objective function; and (ii) updating the surrogate function to tighten the upper bound.

As mentioned in the Introduction, the MISO method proposed by Mairal [2015] is developed as an iterative scheme that only updates the surrogate functions *partially* at each iteration. Formally, for any $i \in [\![1, n]\!]$, we consider a surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ which satisfies

S1. *For all $i \in [\![1, n]\!]$ and $\overline{\boldsymbol{\theta}} \in \Theta$, the function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ is convex* w.r.t. *$\boldsymbol{\theta}$, and it holds*

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta}), \; \forall \, \boldsymbol{\theta} \in \Theta \,, \tag{3}$$

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

S2. *For any $\overline{\boldsymbol{\theta}}_i \in \Theta$, $i \in [\![1, n]\!]$ and some $\epsilon > 0$, the difference function $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}_i) - \mathcal{L}(\boldsymbol{\theta})$ is defined for all $\boldsymbol{\theta} \in \Theta_\epsilon$ and differentiable for all $\boldsymbol{\theta} \in \Theta$, where*

$\Theta_\epsilon = \{\boldsymbol{\theta} \in \mathbb{R}^d, \inf_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \epsilon\}$ *is an* ϵ*-neighborhood set of* Θ*. Moreover, for some constant* L*, the gradient satisfies*

$$\|\nabla \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)\|^2 \leq 2L\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n),\ \forall\, \boldsymbol{\theta} \in \Theta\ . \tag{4}$$

S1 is a common condition used for surrogate optimization, see [Mairal, 2015, Section 2.3]. Meanwhile, S2 can be satisfied when the difference function $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)$ is L-smooth for all $\boldsymbol{\theta} \in \mathbb{R}^d$, where the condition can be implied through applying [Razaviyayn et al., 2013, Proposition 1].

The inequality (3) implies $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta}) > -\infty$ for any $\boldsymbol{\theta} \in \Theta$. The MISO method is an incremental version of the MM method, as summarized by Algorithm 1. As seen in the pseudo code, the MISO method maintains an iteratively updated set of surrogate upper-bound functions $\{\mathcal{A}_i^k(\boldsymbol{\theta})\}_{i=1}^n$ and updates the iterate through minimizing the average of the surrogate functions.

Algorithm 1 MISO method [Mairal, 2015]

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

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

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

Particularly, only one out of the n surrogate functions is updated at each iteration [cf. Line 5] and the sum function $\frac{1}{n}\sum_{i=1}^n \mathcal{A}_i^{k+1}(\boldsymbol{\theta})$ is designed to be 'easy to optimize', for example, it can be a sum of quadratic functions. As such, the MISO method is suitable for large-scale optimization as the computation cost per iteration is independent of n. Moreover, under S1, S2, it was shown that the MISO method converges almost surely to a stationary point of (1) [Mairal, 2015, Proposition 3.1].

We now consider the case when the surrogate functions $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ are intractable. Let Z be a measurable set, $p_i : \mathsf{Z} \times \Theta \to \mathbb{R}_+$ be a pdf, $r_i : \Theta \times \Theta \times \mathsf{Z} \to \mathbb{R}$ be a measurable function and μ_i be a σ-finite measure, we consider surrogate functions which satisfy S1, S2 that can be expressed as an expectation:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \int_{\mathsf{Z}} r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i) p_i(z_i; \overline{\boldsymbol{\theta}}) \mu_i(dz_i) \quad \forall\, (\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) \in \Theta \times \Theta\ . \tag{5}$$

Plugging (5) into the MISO method is not feasible since the update step in Step 6 involves a minimization of an expectation. Several motivating examples of (1) are given in Section 2.

We propose the *Minimization by Incremental Stochastic Surrogate Optimization* (MISSO) method which replaces the expectation in (5) by *Monte Carlo* integration and then optimizes (1) incrementally. Denote by $M \in \mathbb{N}$ the Monte Carlo batch size and let $z_m \in \mathsf{Z}$, $m = 1, ..., M$ be a set of samples. These samples can be drawn (Case 1) i.i.d. from the distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$ or (Case 2) from a Markov chain with the stationary distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$; see Section 3 for illustrations. To this end, we define

$$\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M) := \frac{1}{M} \sum_{m=1}^M r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_m) \tag{6}$$

and we summarize the proposed MISSO method in Algorithm 2. As seen, the method is similar to the MISO method but it involves two types of randomness. The first randomness comes from the selection of i_k in Line 5. The second randomness is that a set of Monte-Carlo approximated functions $\widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta})$ is used in lieu of $\mathcal{A}_i^k(\boldsymbol{\theta})$ when optimizing for the next iterate $\boldsymbol{\theta}^{(k)}$. We now discuss two applications of the MISSO method.

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

$$g_i(\boldsymbol{\theta}) := \int_{\mathsf{Z}} f_i(z_i, \boldsymbol{\theta}) \mu_i(\mathrm{d}z_i),\ i \in [\![1, n]\!] \tag{9}$$

Algorithm 2 MISSO method

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

$$\widetilde{\mathcal{A}}_i^0(\boldsymbol{\theta}) := \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(k)}}),\ i \in [\![1, n]\!] \ . \tag{7}$$

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

$$\widetilde{\mathcal{A}}_i^{k+1}(\boldsymbol{\theta}) = \begin{cases} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}), & \text{if } i = i_k \\ \widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta}), & \text{otherwise.} \end{cases} \tag{8}$$

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

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

Assume without loss of generality that $g_i(\boldsymbol{\theta}) \neq 0$ for all $\boldsymbol{\theta} \in \Theta$, we define by $p_i(z_i, \boldsymbol{\theta}) := f_i(z_i, \boldsymbol{\theta}) / g_i(\boldsymbol{\theta})$ the conditional distribution of the latent variable z_i given the observation y_i. A surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ satisfying S1 can be obtained through writing $f_i(z_i, \boldsymbol{\theta}) = \frac{f_i(z_i, \boldsymbol{\theta})}{p_i(z_i, \overline{\boldsymbol{\theta}})} p_i(z_i, \overline{\boldsymbol{\theta}})$ and applying the Jensen inequality:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) = \int_{\mathsf{Z}} \underbrace{\log\left(p_i(z_i, \overline{\boldsymbol{\theta}}) / f_i(z_i, \boldsymbol{\theta})\right)}_{= r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i)} p_i(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i) \ , \tag{10}$$

We note that S2 can also be verified for common distribution models. We can apply the MISSO method following the above specification of $r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i), p_i(z_i, \overline{\boldsymbol{\theta}})$.

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

$$p(y, w|x) = \pi(w) \prod_{i=1}^n p(y_i|x_i, w) \ . \tag{11}$$

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

$$\min_{\boldsymbol{\theta} \in \Theta} \ \mathcal{L}(\boldsymbol{\theta}) := \mathrm{KL}\left(q(w; \boldsymbol{\theta}) || p(w|y, x)\right) := \mathbb{E}_{q(w; \boldsymbol{\theta})}\left[\log\left(q(w; \boldsymbol{\theta}) / p(w|y, x)\right)\right] \ . \tag{12}$$

Using (11), we decompose $\mathcal{L}(\boldsymbol{\theta}) = n^{-1} \sum_{i=1}^n \mathcal{L}_i(\boldsymbol{\theta}) + \text{const.}$ where:

$$\mathcal{L}_i(\boldsymbol{\theta}) := -\mathbb{E}_{q(w; \boldsymbol{\theta})}\left[\log p(y_i|x_i, w)\right] + \frac{1}{n} \mathbb{E}_{q(w; \boldsymbol{\theta})}\left[\log q(w; \boldsymbol{\theta}) / \pi(w)\right] = r_i(\boldsymbol{\theta}) + d(\boldsymbol{\theta}) \ . \tag{13}$$

Directly optimizing the finite sum objective function in (12) can be difficult. First, with $n \gg 1$, evaluating the objective function $\mathcal{L}(\boldsymbol{\theta})$ requires a full pass over the entire dataset. Second, for some complex models, the expectations in (13) can be intractable even if we assume a simple parametric model for $q(w; \boldsymbol{\theta})$. Assume that $\mathcal{L}_i$ is L-smooth, *i.e.,* $\mathcal{L}_i$ is differentiable on Θ and its gradient $\nabla \mathcal{L}_i$ is L-Lipschitz. We apply the MISSO method with a quadratic surrogate function defined as:

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \mathcal{L}_i(\overline{\boldsymbol{\theta}}) + \left\langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_i(\overline{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{\mathrm{L}}{2} \|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|^2 \ . \tag{14}$$

It is easily checked that $\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}})$ satisfies S1, S2. To compute the gradient $\nabla\mathcal{L}_i(\overline{\boldsymbol{\theta}})$, we apply the re-parametrization technique suggested in [Paisley et al., 2012, Kingma and Welling, 2014, Blundell et al., 2015]. Let $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$ be a differentiable function *w.r.t.* $\boldsymbol{\theta} \in \Theta$ which is designed such that the law of $w = t(z, \overline{\boldsymbol{\theta}})$, where $z \sim \mathcal{N}_d(0, \mathbf{I})$, is $q(\cdot, \overline{\boldsymbol{\theta}})$. By [Blundell et al., 2015, Proposition 1], the gradient of $-r_i(\cdot)$ in (13) is:

$$\nabla_{\boldsymbol{\theta}}\mathbb{E}_{q(w;\overline{\boldsymbol{\theta}})}\left[\log p(y_i|x_i, w)\right] = \mathbb{E}_{z\sim\mathcal{N}_d(0,\mathbf{I})}\left[\mathrm{J}^t_{\boldsymbol{\theta}}(z, \overline{\boldsymbol{\theta}})\nabla_w \log p(y_i|x_i, w)\big|_{w=t(z,\overline{\boldsymbol{\theta}})}\right] , \tag{15}$$

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

$$r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z) := \left\langle \nabla_{\boldsymbol{\theta}} d(\overline{\boldsymbol{\theta}}) - \mathrm{J}^t_{\boldsymbol{\theta}}(z, \overline{\boldsymbol{\theta}})\nabla_w \log p(y_i|x_i, w)\big|_{w=t(z,\overline{\boldsymbol{\theta}})} \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{L}{2}\|\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}\|^2 . \tag{16}$$

Finally, using (14) and (16), the surrogate function (6) is given by $\widetilde{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M) := M^{-1}\sum_{m=1}^M r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z_m)$ where $\{z_m\}_{m=1}^M$ is an i.i.d sample from $\mathcal{N}(0, \mathbf{I})$.

3 Convergence Analysis

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

H1. *For all $i \in [\![1, n]\!]$, $\overline{\boldsymbol{\theta}} \in \Theta$, $z_i \in \mathsf{Z}$, the measurable function $r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z_i)$ is convex in $\boldsymbol{\theta}$ and is lower bounded.*

H2. *For the samples $\{z_{i,m}\}_{m=1}^M$, there exists finite constants C_{r} and C_{gr} such that*

$$C_{\mathsf{r}} := \sup_{\overline{\boldsymbol{\theta}}\in\Theta}\sup_{M>0}\frac{1}{\sqrt{M}}\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|\sum_{m=1}^M\left\{r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}, z_{i,m}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}})\right\}\right|\right] \tag{17}$$

$$C_{\mathsf{gr}} := \sup_{\overline{\boldsymbol{\theta}}\in\Theta}\sup_{M>0}\sqrt{M}\mathbb{E}_{\overline{\boldsymbol{\theta}}}\left[\sup_{\boldsymbol{\theta}\in\Theta}\left|\frac{1}{M}\sum_{m=1}^M\frac{\widehat{\mathcal{L}}'_i(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}};\overline{\boldsymbol{\theta}}) - r'_i(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}};\overline{\boldsymbol{\theta}}, z_{i,m})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|}\right|^2\right] \tag{18}$$

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

H2 essentially requires to control the expectation of the supremum of an empirical process [Shapiro et al., 2009, Boucheron et al., 2013]. In particular, if $M \to \infty$, the surrogate function's value and its directional derivative approximate that of $\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}})$ uniformly for all $\boldsymbol{\theta} \in \Theta$. As discussed before, there are two relevant cases here:

Case 1: When the samples $\{z_m\}_{m=1}^M$ used to construct the approximation $\widetilde{\mathcal{L}}_i(\cdot;\cdot,)$ are drawn i.i.d. directly from $p_i(\cdot;\overline{\boldsymbol{\theta}})$ and Θ is bounded, then H2 can be implied by the concentration of measure under certain additional regularity conditions.

Case 2: When the samples are generated by an MCMC procedure, H2 can be achieved through an maximal inequality for beta-mixing sequences obtained in [Doukhan et al., 1995]. The condition may also be implied by a number of drift and minorization conditions [Meyn and Tweedie, 2012].

Stationarity measure As problem (1) is a constrained optimization, we consider the following stationarity measure:

$$g(\overline{\boldsymbol{\theta}}) := \inf_{\boldsymbol{\theta}\in\Theta}\frac{\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \quad \text{and} \quad g(\overline{\boldsymbol{\theta}}) = g_+(\overline{\boldsymbol{\theta}}) - g_-(\overline{\boldsymbol{\theta}}) , \tag{19}$$

where $g_+(\overline{\boldsymbol{\theta}}) := \max\{0, g(\overline{\boldsymbol{\theta}})\}$, $g_-(\overline{\boldsymbol{\theta}}) := -\min\{0, g(\overline{\boldsymbol{\theta}})\}$ denote the positive and negative part of $g(\overline{\boldsymbol{\theta}})$, respectively. Note that $\overline{\boldsymbol{\theta}}$ is a stationary point if and only if $g_-(\overline{\boldsymbol{\theta}}) = 0$ [Fletcher et al., 2002]. Furthermore, suppose that the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ has a limit point $\overline{\boldsymbol{\theta}}$ that is a stationary point, then one has $\lim_{k\to\infty} g_-(\boldsymbol{\theta}^{(k)}) = 0$. In this sense, the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\geq 0}$ is said to satisfy an *asymptotic stationary point condition*. This is equivalent to [Mairal, 2015, Definition 2.4].

To explain the condition (19), observe that if $\overline{\boldsymbol{\theta}} \in \mathrm{int}(\Theta)$, the directional derivative can be replaced by the inner product between the gradient $\nabla\mathcal{L}(\overline{\boldsymbol{\theta}})$ and $\boldsymbol{\theta} - \overline{\boldsymbol{\theta}}$, *i.e.,* $\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}) = \langle \nabla\mathcal{L}(\overline{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \rangle$. Therefore, from the definition we have $g(\overline{\boldsymbol{\theta}}) = -\|\nabla\mathcal{L}(\overline{\boldsymbol{\theta}})\| = -g_-(\overline{\boldsymbol{\theta}})$. If in addition $g_-(\overline{\boldsymbol{\theta}}) = 0$, then $\overline{\boldsymbol{\theta}}$ is a stationary point to (1) in the same sense as in unconstrained optimization.

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

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n}\textstyle\sum_{i=1}^n \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{20}$$

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

Theorem 1. *Under S1, S2, H1, H2. For any $K_{\mathsf{max}} \in \mathbb{N}$, let K be an independent discrete r.v. drawn uniformly from $\{0, ..., K_{\mathsf{max}} - 1\}$ and define the following quantity:*

$$\Delta_{(K_{\mathsf{max}})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})] + \sum_{k=0}^{K_{\mathsf{max}}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}}\,, \tag{21}$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}\big[\|\nabla\widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] \le \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}, \;\; \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \le \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}. \tag{22}$$

Note that $\Delta_{(K_{\mathsf{max}})}$ is finite for any $K_{\mathsf{max}} \in \mathbb{N}$. As expected, the MISSO method converges to a stationary point of (1) asymptotically and at a sublinear rate $\mathbb{E}[g_-^{(K)}] \le -\mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$. Furthermore, we remark that the MISO method can be analyzed in Theorem 1 as a special case of the MISSO method satisfying $C_{\mathsf{r}} = C_{\mathsf{gr}} = 0$. In this case, while the asymptotic convergence is well known from [Mairal, 2015] [cf. H2], Eq. (22) gives a non-asymptotic rate of $\mathbb{E}[g_-^{(K)}] \le -\mathcal{O}(\sqrt{nL/K_{\mathsf{max}}})$ which is new to our best knowledge.

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

Theorem 2. *Under S1, S2, H1, H2. In addition, assume that $\{M_{(k)}\}_{k\ge0}$ is a non-decreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$. Then:*

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

In particular, the first result above shows that the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k\ge0}$ produced by the MISSO method satisfies an *asymptotic stationary point condition*.

4 Numerical Experiments

4.1 Binary logistic regression with missing values

This application follows **Example 1** described in Section 2. We consider a binary regression setup, $((y_i, z_i), i \in [\![n]\!])$ where $y_i \in \{0, 1\}$ is a binary response and $z_i = (z_{i,j} \in \mathbb{R}, j \in [\![p]\!])$ is a covariate vector. The vector of covariates z_i is not fully observed. We denote by $z_{i,mis}$ the missing values and $z_{i,obs}$ the observed covariate. It is assumed that $(z_i, i \in [\![n]\!])$ are i.i.d. and marginally distributed according to $\mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Omega})$ where $\beta \in \mathbb{R}^p$ and Ω is a positive semidefinite $p \times p$ matrix. Here, $\boldsymbol{\theta} = (\boldsymbol{\beta}, \boldsymbol{\Omega})$ is the parameter to be estimated.

We are interested in finding the latent structure of the covariates z_i. We define the conditional distribution of the observations y_i given $z_i = (z_{i,mis}, z_{i,obs})$ as:

$$p_i(y_i|z_i) = \frac{\exp(-y_i\delta^\top \bar{z}_i)}{1 + \exp(-\delta^\top \bar{z}_i)}\,, \tag{23}$$

where $\delta = (\delta_0, \cdots, \delta_p)$ are, for simplicity, assumed to be known and $\bar{z}_i = (1, z_i)$. For $i \in [\![n]\!]$, the complete data log-likelihood is expressed as:

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

This model belongs to the curved exponential family [Keener, 2010] where for all $i \in [\![n]\!]$, the complete data sufficient statistics are given by $\tilde{S}_i(z_i) \triangleq (z_i, z_i z_i^\top)$, see Appendix D.1.2.

At the k-th iteration, and after the initialization, for all $i \in [\![n]\!]$, of the statistics $(s_i^{1,(0)}, s_i^{2,(0)})$, the MISSO algorithm consists in picking an index i_k uniformly on $[\![n]\!]$, sampling a Monte Carlo batch $\{z_{i_k,mis,m}^{(k)}\}_{m=1}^{M_{(k)}}$ from the conditional distribution $p(z_{i_k,mis}|z_{i_k,obs}, y_{i_k}; \boldsymbol{\theta}^{(k-1)})$ using an MCMC sampler and computing the quantities $(s_i^{1,(k)}, s_i^{2,(k)})$ as follows:

$$(s_i^{1,(k)}, s_i^{2,(k)}) = \begin{cases} \left(\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} z_{i,m}^{(k)}, \frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} z_{i,m}^{(k)}(z_{i,m}^{(k)})^\top\right) & \text{if } i = i_k \\ (s_i^{1,(k-1)}, s_i^{2,(k-1)}) & \text{otherwise} \end{cases} \tag{24}$$

where $z_{i,m}^{(k)} = (z_{i,mis,m}^{(k)}, z_{i,obs})$ is composed of a simulated and an observed part. Finally, set

$$\boldsymbol{\beta}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} s_i^{1,(k)} \quad \text{and} \quad \boldsymbol{\Omega}^{(k)} = \frac{1}{n}\sum_{i=1}^{n} s_i^{2,(k)} - \boldsymbol{\beta}^{(k)}(\boldsymbol{\beta}^{(k)})^\top. \tag{25}$$

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

Similar to [Jiang et al., 2018], we select $p = 16$ influential quantitative measurements, described in Appendix D.1.1, on $n = 6384$ patients, and we adopt the logistic regression model with missing covariates in (23) to predict the risk of a severe hemorrhage which is one of the main cause of death after a major trauma. Note as the dataset considered is heterogeneous – coming from multiple sources with frequently missed entries – we apply the latent data model described in the above. For the Monte-Carlo sampling of $z_{i,mis}$, we run a Metropolis Hastings algorithm with the target distribution $p(\cdot|z_{i,obs}, y_i; \boldsymbol{\theta}^{(k)})$ whose procedure is detailed in Appendix D.1.3.

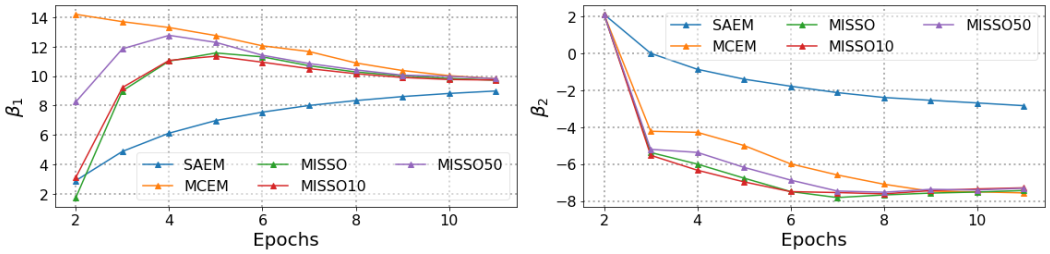

Figure 1: Convergence of two components of the vector of parameters β for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against the number of passes over the data.

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

4.2 Fitting Bayesian LeNet-5 on MNIST

This application follows **Example 2** described in Section 2. We apply the MISSO method to fit a Bayesian variant of LeNet-5 [LeCun et al., 1998] (see Appendix D.2.1). We train this network on

the MNIST dataset [LeCun, 1998]. The training set is composed of $N = 55\,000$ handwritten digits, 28×28 images. Each image is labelled with its corresponding number (from zero to nine). Under the prior distribution π, see (11), the weights are assumed independent and identically distributed according to $\mathcal{N}(0, 1)$. We also assume that $q(\cdot; \boldsymbol{\theta}) \equiv \mathcal{N}(\mu, \sigma^2 \mathbf{I})$. The variational posterior parameters are thus $\boldsymbol{\theta} = (\mu, \sigma)$ where $\mu = (\mu_\ell, \ell \in [\![d]\!])$ where d is the number of weights in the neural network. We use the re-parametrization as $w = t(\boldsymbol{\theta}, z) = \mu + \sigma z$ with $z \sim \mathcal{N}(0, \mathbf{I})$.

At iteration k, minimizing the sum of stochastic surrogates defined as in (6) and (16) yields the following MISSO update — step (i) pick a function index i_k uniformly on $[\![n]\!]$; step (ii) sample a Monte Carlo batch $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ from $\mathcal{N}(0, \mathbf{I})$; and step (iii) update the parameters as

$$\mu_\ell^{(k)} = \frac{1}{n}\sum_{i=1}^n \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^n \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n}\sum_{i=1}^n \sigma^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^n \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} , \tag{26}$$

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

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

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

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

We compare the convergence of the *Monte Carlo variants* of the following state of the art optimization algorithms — the ADAM [Kingma and Ba, 2015], the Momentum [Sutskever et al., 2013] and the SAG [Schmidt et al., 2017] methods versus the *Bayes by Backprop* (BBB) [Blundell et al., 2015] and our proposed MISSO method. For all these methods, the loss function (13) and its gradients were computed by Monte Carlo integration using Tensorflow Probability library [Dillon et al., 2017], based on the re-parametrization described above. Update rules for each algorithm are performed using their vanilla implementations on TensorFlow [Abadi et al., 2015] as detailed in Appendix D.2.2. We use the following hyperparameters for all runs — the learning rate is 10^{-3}, we run 100 epochs with a mini-batch size of 128 and use the batchsize of $M_{(k)} = k$.

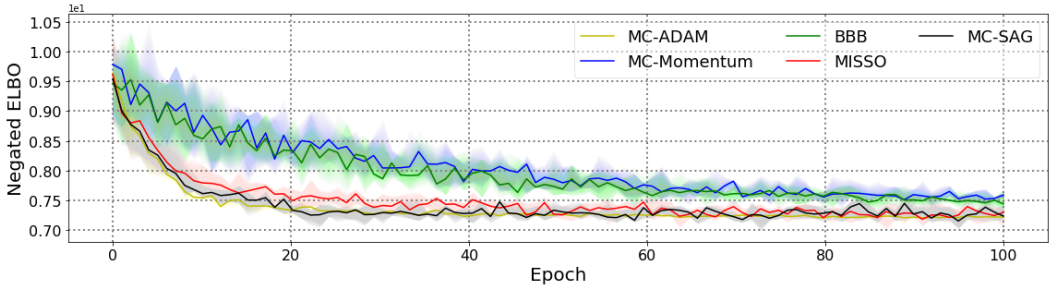

Figure 2: (Incremental Variational Inference) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

Figure 2 shows the convergence of the negated evidence lower bound against the number of passes over data (one pass represents an epoch). As observed, the proposed MISSO method outperforms *Bayes by Backprop* and Momentum, while similar convergence rates are observed with the MISSO, ADAM and SAG methods.

5 Conclusions

We present a unifying framework for minimizing a non-convex finite-sum objective function using incremental surrogates when the latter functions are expressed as an expectation and are intractable.

Our approach covers a large class of non-convex applications in machine learning such as logistic regression with missing values and variational inference. We provide both finite-time and asymptotic guarantees of our incremental stochastic surrogate optimization technique and illustrate our findings training a binary logistic regression with missing covariates to predict hemorrhagic shock and a Bayesian variant of LeNet-5 on MNIST.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G. 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.

C. M. Bishop. *Pattern recognition and machine learning*. springer, 2006.

D. M. Blei, A. Kucukelbir, and J. D. McAuliffe. Variational inference: A review for statisticians. *Journal of the American Statistical Association*, 112(518):859–877, 2017. doi: 10.1080/01621459.2017.1285773. URL https://doi.org/10.1080/01621459.2017.1285773.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.

S. Boucheron, G. Lugosi, and P. Massart. *Concentration inequalities: A nonasymptotic theory of independence*. Oxford university press, 2013.

B. Delyon, M. Lavielle, and E. Moulines. Convergence of a stochastic approximation version of the em algorithm. *Ann. Statist.*, 27(1):94–128, 03 1999. doi: 10.1214/aos/1018031103. URL https://doi.org/10.1214/aos/1018031103.

J. V. Dillon, I. Langmore, D. Tran, E. Brevdo, S. Vasudevan, D. Moore, B. Patton, A. Alemi, M. D. Hoffman, and R. A. Saurous. Tensorflow distributions. *CoRR*, abs/1711.10604, 2017. URL http://arxiv.org/abs/1711.10604.

P. Doukhan, P. Massart, and E. Rio. Invariance principles for absolutely regular empirical processes. In *Annales de l'IHP Probabilités et statistiques*, volume 31, pages 393–427, 1995.

R. Fletcher, N. I. Gould, S. Leyffer, P. L. Toint, and A. Wächter. Global convergence of a trust-region sqp-filter algorithm for general nonlinear programming. *SIAM Journal on Optimization*, 13(3):635–659, 2002.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521(7553):452–459, May 2015. doi: 10.1038/nature14541. URL https://www.ncbi.nlm.nih.gov/pubmed/26017444/. On Probabilistic models.

W. Jiang, J. Josse, and M. Lavielle. Logistic regression with missing covariates–parameter estimation, model selection and prediction. 2018.

M. I. Jordan, Z. Ghahramani, T. S. Jaakkola, and L. K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, Nov. 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL https://doi.org/10.1023/A:1007665907178.

R. Keener. *Curved Exponential Families*, pages 85–99. Springer New York, New York, NY, 2010. ISBN 978-0-387-93839-4.

D. P. Kingma and J. Ba. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015. URL http://arxiv.org/abs/1412.6980.

D. P. Kingma and M. Welling. Auto-encoding variational bayes. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014. URL http://arxiv.org/abs/1312.6114.

K. Lange. *MM Optimization Algorithms*. SIAM-Society for Industrial and Applied Mathematics, USA, 2016. ISBN 1611974399, 9781611974393.

Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Y. Li and Y. Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2052–2061. JMLR. org, 2017.

J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM J. Optim.*, 25(2):829–855, 2015. ISSN 1052-6234. doi: 10.1137/140957639. URL https://doi.org/10.1137/140957639.

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. ISBN 978-0-471-20170-0. doi: 10.1002/9780470191613. URL https://doi.org/10.1002/9780470191613.

S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.

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

J. Paisley, D. Blei, and M. Jordan. Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress, 2012.

N. G. Polson, V. Sokolov, et al. Deep learning: a bayesian perspective. *Bayesian Analysis*, 12(4): 1275–1304, 2017.

M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.

M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.

A. Shapiro, D. Dentcheva, and A. Ruszczyński. *Lectures on stochastic programming: modeling and theory*. SIAM, 2009.

I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.

G. C. G. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association*, 85(411): 699–704, 1990. doi: 10.1080/01621459.1990.10474930. URL https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474930.

A Proof of Theorem 1

Theorem. *Under S1, S2, H1, H2. For any $K_{\mathsf{max}} \in \mathbb{N}$, let K be an independent discrete r.v. drawn uniformly from $\{0, ..., K_{\mathsf{max}} - 1\}$ and define the following quantity:*

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

Then we have following non-asymptotic bounds:

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

Proof We begin by recalling the definition

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta}). \tag{27}$$

Notice that

$$\begin{aligned}\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^{k+1})}, \{z_{i,m}^{(\tau_i^{k+1})}\}_{m=1}^{M_{(\tau_i^{k+1})}}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big).\end{aligned} \tag{28}$$

Furthermore, we recall that

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n}\textstyle\sum_{i=1}^{n}\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{29}$$

Due to S2, we have

$$\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{30}$$

To prove the first bound in (22), using the optimality of $\boldsymbol{\theta}^{(k+1)}$, one has

$$\begin{aligned}&\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big)\end{aligned} \tag{31}$$

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

$$\begin{aligned}&\mathbb{E}_{i_k}\big[\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}})|\mathcal{F}_k, i_k\big]|\mathcal{F}_k\big] \\ &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathbb{E}_{i_k}\big[\mathbb{E}\big[\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} r_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, z_{i_k,m}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})|\mathcal{F}_k, i_k\big]|\mathcal{F}_k\big] \\ &\leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}},\end{aligned} \tag{32}$$

where the last inequality is due to H2. Moreover,

$$\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})|\mathcal{F}_k\big] = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, \{z_{i,m}^{(\tau_i^k)}\}_{m=1}^{M_{(\tau_i^k)}}) = \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{33}$$

Taking the conditional expectations on both sides of (31) and re-arranging terms give:

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} \tag{34}$$

Proceeding from (34), we observe the following lower bound for the left hand side

$$\begin{aligned}
&\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \overset{(a)}{=} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \\
&\overset{(b)}{\geq} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2L}\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} r_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)})\Big\}}_{:=-\delta^{(k)}(\boldsymbol{\theta}^{(k)})} + \frac{1}{2L}\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2
\end{aligned} \tag{35}$$

where (a) is due to $\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$ [cf. S1], (b) is due to (30) and we have defined the summation in the last equality as $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$. Substituting the above into (34) yields

$$\frac{\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)}) \tag{36}$$

Observe the following upper bound on the total expectations:

$$\mathbb{E}\big[\delta^{(k)}(\boldsymbol{\theta}^{(k)})\big] \leq \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{C_{\mathsf{r}}}{\sqrt{M_{(\tau_i^k)}}}\Big], \tag{37}$$

which is due to H2. It yields

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

Finally, for any $K_{\mathsf{max}} \in \mathbb{N}$, we let K be a discrete r.v. that is uniformly drawn from $\{0, 1, ..., K_{\mathsf{max}} - 1\}$. Using H2 and taking total expectations lead to

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] = \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{2LC_{\mathsf{r}}}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}\Big[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{M_{(\tau_i^k)}}}\Big]
\end{aligned} \tag{38}$$

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

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^{k}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} \tag{39}$$

Taking the sum yields:

$$\begin{aligned}
&\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{j=1}^{k}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{l=0}^{k-1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{k-(l+1)} M_{(l)}^{-1/2} \\
&= \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}\sum_{k=l+1}^{K_{\mathsf{max}}-1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{k-(l+1)} \leq \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}
\end{aligned} \tag{40}$$

where the last inequality is due to upper bounding the geometric series. Plugging this back into (38) yields

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] = \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}} = \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}.
\end{aligned} \tag{41}$$

This concludes our proof for the first inequality in (22).

To prove the second inequality of (22), we define the shorthand notations $g^{(k)} := g(\boldsymbol{\theta}^{(k)})$, $g_-^{(k)} := -\min\{0, g^{(k)}\}$, $g_+^{(k)} := \max\{0, g^{(k)}\}$. We observe that

$$\begin{aligned} g^{(k)} &= \inf_{\boldsymbol{\theta}\in\Theta} \frac{\mathcal{L}'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \\ &= \inf_{\boldsymbol{\theta}\in\Theta} \left\{ \frac{\frac{1}{n}\sum_{i=1}^n \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} - \frac{\left\langle \nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \,|\, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)} \right\rangle}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \right\} \\ &\geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta} \frac{\frac{1}{n}\sum_{i=1}^n \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \end{aligned} \tag{42}$$

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

$$\begin{aligned} &\frac{1}{n}\sum_{i=1}^n \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\ &= \underbrace{\widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})}_{\geq 0} - \widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}) + \frac{1}{n}\sum_{i=1}^n \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\ &\geq \frac{1}{n}\sum_{i=1}^n \left\{ \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) - \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) \right\} \end{aligned} \tag{43}$$

where the inequality is due to the optimality of $\boldsymbol{\theta}^{(k)}$ and the convexity of $\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta})$ [cf. H1]. Denoting a scaled version of the above term as:

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

We have

$$g^{(k)} \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta}(-\epsilon^{(k)}(\boldsymbol{\theta})) \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| - \sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{44}$$

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

$$g_-^{(k)} \leq \|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{45}$$

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

$$\mathbb{E}[g_-^{(K)}] \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|] + \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(K)}(\boldsymbol{\theta})] \tag{46}$$

We note that

$$\left(\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|]\right)^2 \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta(K_{\mathsf{max}})}{K_{\mathsf{max}}}, \tag{47}$$

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

$$\begin{aligned} \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(K)}(\boldsymbol{\theta})] = \frac{1}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}} \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(k)}(\boldsymbol{\theta})] &\overset{(a)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^n M_{(\tau_i^k)}^{-1/2}\Big] \\ &\overset{(b)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2} \end{aligned} \tag{48}$$

where (a) is due to H2 and (b) is due to (40). This implies

$$\mathbb{E}[g_-^{(K)}] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}, \tag{49}$$

and concludes the proof of the theorem. □

B Proof of Theorem 2

Theorem. *Under S1, S2, H1, H2. In addition, assume that $\{M_{(k)}\}_{k\geq 0}$ is a non-decreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$. Then:*

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

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

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

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1} \tag{50}$$

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 $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.

We proceed from (31) by re-arranging terms and observing that

$$\begin{aligned}
\widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq\ & \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tfrac{1}{n}\big(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\big) \\
& - \big(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})\big) + \big(\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\big) \\
& + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\big) \\
& + \tfrac{1}{n}\big(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big)
\end{aligned} \tag{51}$$

Our idea is to apply Lemma 1. Under S1, the finite sum of surrogate functions $\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta})$, defined in (20), is lower bounded by a constant $c_k > -\infty$ for any $\boldsymbol{\theta}$. To this end, we observe that

$$V_k := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \inf_{k\geq 0} c_k \geq 0 \tag{52}$$

is a non-negative random variable.

Secondly, under H1, the following random variable is non-negative

$$X_k := \tfrac{1}{n}\big(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(\tau_{i_k}^k)};\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\big) \geq 0. \tag{53}$$

Thirdly, we define

$$\begin{aligned}
E_k =\ & -\big(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})\big) + \big(\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\big) \\
& + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\big) \\
& + \tfrac{1}{n}\big(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big).
\end{aligned} \tag{54}$$

Note that from the definitions (52), (53), (54), we have $V_{k+1} \leq V_k - X_k + E_k$ for any $k \geq 1$.

Under H2, we observe that

$$\mathbb{E}\big[|\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})|\big] \leq C_{\mathrm{r}} M_{(k)}^{-1/2} \tag{55}$$

$$\mathbb{E}\Big[\Big|\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\Big|\Big] \leq C_{\mathsf{r}}\mathbb{E}\Big[M_{(\tau_{i_k}^k)}^{-1/2}\Big] \tag{56}$$

$$\mathbb{E}\big[|\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})|\big] \leq \tfrac{1}{n}\textstyle\sum_{i=1}^n C_{\mathsf{r}}\mathbb{E}\Big[M_{(\tau_i^k)}^{-1/2}\Big] \tag{57}$$

Therefore,

$$\mathbb{E}\big[|E_k|\big] \leq \tfrac{C_{\mathsf{r}}}{n}\Big(M_{(k)}^{-1/2} + \mathbb{E}\Big[M_{(\tau_{i_k}^k)}^{-1/2} + \textstyle\sum_{i=1}^n\big\{M_{(\tau_i^k)}^{-1/2} + M_{(\tau_i^{k+1})}^{-1/2}\big\}\Big]\Big) \tag{58}$$

Using (40) and the assumption on the sequence $\{M_{(k)}\}_{k\geq 0}$, we obtain that

$$\sum_{k=0}^{\infty}\mathbb{E}\big[|E_k|\big] < \frac{C_{\mathsf{r}}}{n}(2+2n)\sum_{k=0}^{\infty}M_{(k)}^{-1/2} < \infty. \tag{59}$$

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

$$\begin{aligned}\infty > \sum_{k=0}^{\infty}\mathbb{E}[X_k] &= \frac{1}{n}\sum_{k=0}^{\infty}\mathbb{E}\big[\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})\big]\\ &= \frac{1}{n}\sum_{k=0}^{\infty}\mathbb{E}\big[\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})\big] = \frac{1}{n}\sum_{k=0}^{\infty}\mathbb{E}\big[\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\big]\end{aligned} \tag{60}$$

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

$$\lim_{k\to\infty}\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0 \quad \text{a.s.} \tag{61}$$

and subsequently applying (30), we have $\lim_{k\to\infty}\|\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| = 0$ almost surely. Finally, it follows from (30) and (45) that

$$\lim_{k\to\infty} g_-^{(k)} \leq \lim_{k\to\infty}\sqrt{2L}\sqrt{\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})} + \lim_{k\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})| = 0, \tag{62}$$

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

Finally, we prove that $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges almost surely. As a consequence of Lemma 1, it is clear that $\{V_k\}_{k\geq 0}$ converges almost surely and so is $\{\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\}_{k\geq 0}$, *i.e.,* we have $\lim_{k\to\infty}\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$. Applying (61) implies that

$$\underline{\mathcal{L}} = \lim_{k\to\infty}\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \lim_{k\to\infty}\mathcal{L}(\boldsymbol{\theta}^{(k)}) \quad \text{a.s.} \tag{63}$$

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

C Proof of Lemma 1

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

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

then:

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

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \tag{64}$$

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:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \tag{65}$$

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

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \end{aligned} \tag{66}$$

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:

$$\mathbb{E}[\liminf_{k\to\infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k\to\infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \tag{67}$$

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:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k\to\infty} U_k] \tag{68}$$

Finally, we have:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k\to\infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \tag{69}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (66) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\begin{aligned} \sum_{j=1}^{\infty} X_j &\leq W_0 - W_\infty < \infty \\ \sum_{j=1}^{\infty} \mathbb{E}[X_j] &\leq \mathbb{E}[W_0] - w_\infty < \infty \end{aligned} \tag{70}$$

which concludes the proof of the lemma. □

D Details about the Numerical Experiments

D.1 Binary Logistic Regression on the Traumabase

D.1.1 Traumabase quantitative variables

The list of the 16 quantitative variables we use in our experiments are as follows — *age, weight, height, BMI (Body Mass Index), the Glasgow Coma Scale, the Glasgow Coma Scale motor component, the minimum systolic blood pressure, the minimum diastolic blood pressure, the maximum*

number of heart rate (or pulse) per unit time (usually a minute), the systolic blood pressure at arrival of ambulance, the diastolic blood pressure at arrival of ambulance, the heart rate at arrival of ambulance, the capillary Hemoglobin concentration, the oxygen saturation, the fluid expansion colloids, the fluid expansion cristalloids, the pulse pressure for the minimum value of diastolic and systolic blood pressure, the pulse pressure at arrival of ambulance.

D.1.2 MISSO, MCEM and SAEM for Curved Exponential Family

Consider the case where the complete model $\boldsymbol{\theta} \to f_i(z_i, \boldsymbol{\theta})$ belongs to the curved exponential family. For all $i \in [\![1, n]\!]$ and $\boldsymbol{\theta} \in \Theta$, one has:

$$\log f_i(z_i, \boldsymbol{\theta}) = H_i(z_i) - \psi_i(\boldsymbol{\theta}) + \left\langle \tilde{S}_i(z_i) \,|\, \phi_i(\boldsymbol{\theta}) \right\rangle. \tag{71}$$

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

$$\mathcal{L}_i(\boldsymbol{\theta}) = -\log \int_{\mathsf{Z}} f_i(z_i, \boldsymbol{\theta}) \mu_i(\mathrm{d}z_i). \tag{72}$$

To derive the surrogate function for this example, we follow from **Example 1** in Section 2 by taking $p(z_i, \overline{\boldsymbol{\theta}}) = f(z_i, \overline{\boldsymbol{\theta}})/g(\overline{\boldsymbol{\theta}})$ to yield

$$\mathcal{L}_i(\boldsymbol{\theta}) \leq \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \text{constant} + \int_{\mathsf{Z}} \{ -\log f_i(z_i, \boldsymbol{\theta}) \} p(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i) \tag{73}$$

Using (71), the latter integral can be further expressed as

$$\text{constant} + \psi_i(\boldsymbol{\theta}) - \int_{\mathsf{Z}} \left\langle \tilde{S}_i(z_i) \,|\, \phi_i(\boldsymbol{\theta}) \right\rangle p(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i). \tag{74}$$

To summarize, we observe that the surrogate function can be written as $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) = \text{constant} + \psi_i(\boldsymbol{\theta}) - \left\langle \int_{\mathsf{Z}} \tilde{S}_i(z_i) p(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i) \,|\, \phi_i(\boldsymbol{\theta}) \right\rangle$. As suggested in the MISSO method, we note that the integral can be approximated using sample averages.

Now consider the following function $L(s; \boldsymbol{\theta})$:

$$L(s; \boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \left\{ \psi_i(\boldsymbol{\theta}) - \langle s_i \,|\, \phi_i(\boldsymbol{\theta}) \rangle \right\}, \tag{75}$$

for all $\boldsymbol{\theta} \in \Theta$ and $s = (s_i, 1 \leq i \leq N) \in \mathsf{S} := \times_{n=1}^{N} \mathsf{S}_i$. As noted from (74), when s_i is obtained from an appropriately sampling procedure, (75) constitutes the summed stochastic surrogate function $\widetilde{\mathcal{L}}(\boldsymbol{\theta})$ as suggested in the MISSO method [cf. Algorithm 2]. For simplicity, we consider a case when the minimization *w.r.t.* $\boldsymbol{\theta}$ can be uniquely attained and we denote $\boldsymbol{\theta}^{(s)} := \arg\min_{\boldsymbol{\theta} \in \Theta} L(s; \boldsymbol{\theta})$. We note that for this special case, the MISSO method can be compactly described in Algorithm 3.

Now, in the context of the logistic regression described in section 4.1, for all $i \in [\![1, n]\!]$, the complete log likelihood is expressed as:

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

For all $i \in [\![1, n]\!]$, the sufficient statistics is $\tilde{S}_i(z_i) \triangleq (z_i, z_i z_i^\top)$ and the minimizing map $s \to \hat{\boldsymbol{\theta}}(s)$ can be derived as:

$$\hat{\boldsymbol{\theta}}\big((s_{i,1}, s_{i,2})_{i=1}^n\big) = \left(\frac{1}{n} \sum_{i=1}^{n} s_{i,1}, \frac{1}{n} \sum_{i=1}^{n} s_{i,2} - \frac{1}{n^2} \sum_{i=1}^{n} s_{i,1} \Big(\sum_{i=1}^{n} s_{i,1} \Big)^\top \right) \tag{77}$$

Let us compare the MISSO method to the MCEM, SAEM methods as follows. First we note that the MISSO method differs from the MCEM method in that the latter picks *all the data sample* at each iteration, *i.e.,* Line 3 to 5 is performed for every $i \in [\![1, n]\!]$ instead of only one randomly chosen

Algorithm 3 MISSO for curved exponential family

Initialization: given an initial parameter estimate $\boldsymbol{\theta}^{(0)}$, for all $i \in [\![1, n]\!]$, sample a Monte Carlo batch $\{z_{i,m}^{(0)}\}_{m=1}^{M_{(0)}}$ from $p(z_i, \boldsymbol{\theta}^{(0)})$ and compute $s_i^{(0)} = \frac{1}{M_{(0)}} \sum_{m=1}^{M_{(0)}} \tilde{S}_i(z_{i,m}^{(0)})$.
for $k = 0, 1..., $ **do**
  Pick a function index i_k uniformly on $[\![1, n]\!]$.
  Draw $M_{(k)}$ Monte-Carlo samples $\{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}$ from $p(z_i, \boldsymbol{\theta}^{(k-1)})$.
  Update the individual sufficient statistics recursively — for each $i \in [\![1, n]\!]$,

$$s_i^{(k)} = \begin{cases} \frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \tilde{S}_i(z_{i,m}^{(k)}) & \text{if } i = i_k \\ s_i^{(k-1)} & \text{otherwise} \end{cases} \tag{76}$$

  Set $\boldsymbol{\theta}^{(k)} = \boldsymbol{\theta}^{(s^{(k)})} = \arg\min_{\boldsymbol{\theta} \in \Theta} L(s^{(k)}; \boldsymbol{\theta})$ where $s^{(k)} = (s_i^{(k)}, 1 \leq i \leq n)$.
end for

Algorithm 4 MH aglorithm

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

$i_k \in [\![1, n]\!]$. Second, the SAEM method consists of a different update of the individual sufficient statistics — at iteration k and for all $i \in [\![1, n]\!]$, the SAEM updates the quantity $s_i^{(k)}$ as follows:

$$s_i^{(k)} = s_i^{(k-1)} + \gamma_k \left(\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \tilde{S}_i(z_{i,m}^{(k)}) - s_i^{(k-1)} \right) \tag{78}$$

where $\{\gamma_k\}_{k>0}$ is a decreasing sequence of stepsizes. Note that when $\gamma_k = 1$, the above update recovers the MCEM method.

D.1.3 Metropolis Hastings algorithm

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

$$\begin{aligned} \xi &= \beta_{miss} + \Omega_{mis,obs} \Omega_{obs,obs}^{-1} (z_{i,obs} - \beta_{obs}) \ , \\ \Sigma &= \Omega_{mis,mis} + \Omega_{mis,obs} \Omega_{obs,obs}^{-1} \Omega_{obs,mis} \end{aligned} \tag{79}$$

where we have used the Schur Complement of $\Omega_{obs,obs}$ in Ω and noted β_{mis} (resp. β_{obs}) the missing (resp. observed) elements of β. The MH algorithm is summarized in Algorithm 4.

D.2 Incremental Variational Inference for MNIST

D.2.1 Bayesian LeNet-5 Architecture

We describe in Table D.2.1 the architecture of the Convolutional Neural Network introduced in [LeCun et al., 1998] and trained on MNIST:

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

Table 1: LeNet-5 architecture

D.2.2 Algorithms updates

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

$$\begin{aligned}
\hat{\boldsymbol{\delta}}_{\mu_\ell, i_k}^{(k)} &= -\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} \nabla_w \log p(y_{i_k} | x_{i_k}, w) \Big|_{w = t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} d(\boldsymbol{\theta}^{(k-1)}) \, , \\
\hat{\boldsymbol{\delta}}_{\sigma, i_k}^{(k)} &= -\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} z_m^{(k)} \nabla_w \log p(y_{i_k} | x_{i_k}, w) \Big|_{w = t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\sigma} d(\boldsymbol{\theta}^{(k-1)}) \, .
\end{aligned} \tag{80}$$

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

Monte Carlo SAG update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\mu_\ell, i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n} \sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\sigma, i}^{(k)} \, , \tag{81}$$

where $\hat{\boldsymbol{\delta}}_{\mu_\ell, i}^{(k)} = \hat{\boldsymbol{\delta}}_{\mu_\ell, i}^{(k-1)}$ and $\hat{\boldsymbol{\delta}}_{\sigma, i}^{(k)} = \hat{\boldsymbol{\delta}}_{\sigma, i}^{(k-1)}$ for $i \neq i_k$ and are defined by (80) for $i = i_k$. where the learning rate $\gamma = 10^{-3}$.

Bayes By Backprop update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n} \hat{\boldsymbol{\delta}}_{\mu_\ell, i_k}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n} \hat{\boldsymbol{\delta}}_{\sigma, i_k}^{(k)} \, , \tag{82}$$

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

Monte Carlo Momentum update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} + \hat{\boldsymbol{v}}_{\mu_\ell}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} + \hat{\boldsymbol{v}}_{\sigma}^{(k)} \, , \tag{83}$$

where

$$\hat{\boldsymbol{v}}_{\mu_\ell, i}^{(k)} = \alpha \hat{\boldsymbol{v}}_{\mu_\ell}^{(k-1)} - \frac{\gamma}{n} \hat{\boldsymbol{\delta}}_{\mu_\ell, i_k}^{(k)} \quad \text{and} \quad \hat{\boldsymbol{v}}_{\sigma}^{(k)} = \alpha \hat{\boldsymbol{v}}_{\sigma}^{(k-1)} - \frac{\gamma}{n} \hat{\boldsymbol{\delta}}_{\sigma, i_k}^{(k)} \, , \tag{84}$$

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

Monte Carlo ADAM update: Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n} \hat{\boldsymbol{m}}_{\mu_\ell}^{(k)} / (\sqrt{\hat{\boldsymbol{m}}_{\mu_\ell}^{(k)}} + \epsilon) \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n} \hat{\boldsymbol{m}}_{\sigma}^{(k)} / (\sqrt{\hat{\boldsymbol{m}}_{\sigma}^{(k)}} + \epsilon) \, , \tag{85}$$

where

$$\begin{aligned}\hat{\boldsymbol{m}}_{\mu_\ell}^{(k)} &= \boldsymbol{m}_{\mu_\ell}^{(k-1)}/(1-\rho_1^k) \quad \text{with} \quad \boldsymbol{m}_{\mu_\ell}^{(k)} = \rho_1 \boldsymbol{m}_{\mu_\ell}^{(k-1)} + (1-\rho_1)\hat{\boldsymbol{\delta}}_{\mu_\ell,i_k}^{(k)} \,, \\ \hat{\boldsymbol{v}}_{\mu_\ell}^{(k)} &= \boldsymbol{v}_{\mu_\ell}^{(k-1)}/(1-\rho_2^k) \quad \text{with} \quad \boldsymbol{v}_{\mu_\ell}^{(k)} = \rho_2 \boldsymbol{v}_{\mu_\ell}^{(k-1)} + (1-\rho_1)\big(\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)}\big)^2\end{aligned} \tag{86}$$

and

$$\begin{aligned}\hat{\boldsymbol{m}}_{\sigma}^{(k)} &= \boldsymbol{m}_{\sigma}^{(k-1)}/(1-\rho_1^k) \quad \text{with} \quad \boldsymbol{m}_{\sigma}^{(k)} = \rho_1 \boldsymbol{m}_{\sigma}^{(k-1)} + (1-\rho_1)\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)} \,, \\ \hat{\boldsymbol{v}}_{\sigma}^{(k)} &= \boldsymbol{v}_{\sigma}^{(k-1)}/(1-\rho_2^k) \quad \text{with} \quad \boldsymbol{v}_{\sigma}^{(k)} = \rho_2 \boldsymbol{v}_{\sigma}^{(k-1)} + (1-\rho_1)\big(\hat{\boldsymbol{\delta}}_{\sigma,i_k}^{(k)}\big)^2 \,.\end{aligned} \tag{87}$$

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