

---

# MISSO: Minimization by Incremental Stochastic Surrogate Optimization for Large Scale Nonconvex and Nonsmooth Problems

---

Anonymous Author(s)

Affiliation

Address

email

## 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 LeNet and Resnet Bayesian variants on respectively the MNIST and CIFAR-10 datasets.

## 1 Introduction

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

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

where  $\Theta$  is a convex, compact, and closed subset of  $\mathbb{R}^p$ , and for any  $i \in \llbracket 1, n \rrbracket$ , the function  $\mathcal{L}_i : \mathbb{R}^p \rightarrow \mathbb{R}$  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 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]. An extended analysis of MISO in both the convex and nonconvex case has recently been proposed in [Qian et al., 2019].

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.
- We present an incremental update of the commonly used variational inference and Monte-Carlo EM methods as special cases of our newly introduced framework. The analysis of those two algorithms is thus done under this unifying framework of analysis.
- 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  $\epsilon$ -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 intractable surrogate functions requiring approximation. In Section 3, we provide the asymptotic and non-asymptotic convergence analysis for the MISSO method (and of the MISO [Mairal, 2015] one as a special case). Finally, Section 4 presents numerical applications to illustrate our findings including parameter inference for logistic regression with missing covariates and variational inference for two types of Bayesian neural networks.

## TO COMPLETE WITH NOTATIONS

## 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  $i \in \llbracket 1, n \rrbracket$ , we consider a surrogate function  $\hat{\mathcal{L}}_i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}})$  which satisfies

**S1.** For all  $i \in \llbracket 1, n \rrbracket$  and  $\bar{\boldsymbol{\theta}} \in \Theta$ , the function  $\hat{\mathcal{L}}_i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}})$  is convex w.r.t.  $\boldsymbol{\theta}$ , and it holds

$$\hat{\mathcal{L}}_i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta}), \forall \boldsymbol{\theta} \in \Theta, \quad (2)$$

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

**S2.** For any  $\bar{\boldsymbol{\theta}}_i \in \Theta$ ,  $i \in \llbracket 1, n \rrbracket$  and some  $\epsilon > 0$ , the difference function  $\hat{e}(\boldsymbol{\theta}; \{\bar{\boldsymbol{\theta}}_i\}_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \hat{\mathcal{L}}_i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}_i) - \mathcal{L}(\boldsymbol{\theta})$  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  $\epsilon$ -neighborhood set of  $\Theta$ . Moreover, for some constant  $L$ , the gradient satisfies

$$\|\nabla \hat{e}(\boldsymbol{\theta}; \{\bar{\boldsymbol{\theta}}_i\}_{i=1}^n)\|^2 \leq 2L \hat{e}(\boldsymbol{\theta}; \{\bar{\boldsymbol{\theta}}_i\}_{i=1}^n), \forall \boldsymbol{\theta} \in \Theta. \quad (3)$$

**S1** is a common condition used for surrogate optimization, see [Mairal, 2015, Section 2.3]. Meanwhile, **S2** can be satisfied when the difference function  $\hat{e}(\boldsymbol{\theta}; \{\bar{\boldsymbol{\theta}}_i\}_{i=1}^n)$  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 (2) implies  $\widehat{\mathcal{L}}_i(\theta; \bar{\theta}) \geq \mathcal{L}_i(\theta) > -\infty$  for any  $\theta \in \Theta$ . 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  $\{\mathcal{A}_i^k(\theta)\}_{i=1}^n$  and updates the iterate through minimizing the average of the surrogate functions.

Particularly, only one out of the  $n$  surrogate functions is updated at each iteration [cf. Line 5] and the sum function  $\frac{1}{n} \sum_{i=1}^n \mathcal{A}_i^{k+1}(\theta)$  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  $\widehat{\mathcal{L}}_i(\theta; \bar{\theta})$  are intractable. Let  $Z$  be a measurable set,  $p_i : Z \times \Theta \rightarrow \mathbb{R}_+$  be a pdf,  $r_i : \Theta \times \Theta \times Z \rightarrow \mathbb{R}$  be a measurable function and  $\mu_i$  be a  $\sigma$ -finite measure, we consider surrogate functions which satisfy S1, S2 that can be expressed as an expectation:

$$\widehat{\mathcal{L}}_i(\theta; \bar{\theta}) := \int_Z r_i(\theta; \bar{\theta}, z_i) p_i(z_i; \bar{\theta}) \mu_i(dz_i) \quad \forall (\theta, \bar{\theta}) \in \Theta \times \Theta. \quad (4)$$

Plugging (4) 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 (4) by *Monte Carlo* integration and then optimizes (1) incrementally. Denote by  $M \in \mathbb{N}$  the Monte Carlo batch size and let  $z_m \in Z$ ,  $m = 1, \dots, M$  be a set of samples. These samples can be drawn (Case 1) i.i.d. from the distribution  $p_i(\cdot; \bar{\theta})$  or (Case 2) from a Markov chain with the stationary distribution  $p_i(\cdot; \bar{\theta})$ ; see Section 3 for illustrations. To this end, we define

$$\widetilde{\mathcal{L}}_i(\theta; \bar{\theta}, \{z_m\}_{m=1}^M) := \frac{1}{M} \sum_{m=1}^M r_i(\theta; \bar{\theta}, z_m) \quad (5)$$

and we summarize the proposed MISSO method in Algorithm 2. As seen, the procedure is similar to the MISO method but it involves two types of randomness. The first randomness comes from the selection of  $i_k$  in Line 5. The second randomness is that a set of Monte-Carlo approximated functions  $\widetilde{\mathcal{A}}_i^k(\theta)$  is used in lieu of  $\mathcal{A}_i^k(\theta)$  when optimizing for the next iterate  $\theta^{(k)}$ . 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  $((y_i, z_i), i \in \llbracket n \rrbracket)$  where  $y_i$  is observed and  $z_i$  is latent. In this incomplete data framework, define  $\{f_i(z_i, \theta), \theta \in \Theta\}$  to be the complete data likelihood models, i.e., joint likelihood of the observations and latent variables. Let

$$g_i(\theta) := \int_Z f_i(z_i, \theta) \mu_i(dz_i), \quad i \in \llbracket 1, n \rrbracket \quad (8)$$

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  $\mathcal{L}_i(\theta)$  to be the  $i$ th negated incomplete data log-likelihood  $\mathcal{L}_i(\theta) := -\log g_i(\theta)$ .

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

---

#### Algorithm 1 MISO method [Mairal, 2015]

---

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

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

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

---

**Algorithm 2** MISSO method
 

---

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

$$\tilde{\mathcal{A}}_i^0(\theta) := \tilde{\mathcal{L}}_i(\theta; \theta^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(0)}}), \quad i \in \llbracket 1, n \rrbracket. \quad (6)$$

- 4: **for**  $k = 0, 1, \dots$  **do**
- 5:   Pick a function index  $i_k$  uniformly on  $\llbracket 1, n \rrbracket$ .
- 6:   Draw  $M_{(k)}$  Monte-Carlo samples with the stationary distribution  $p_{i_k}(\cdot; \theta^{(k)})$ .
- 7:   Update the individual surrogate functions recursively as:

$$\tilde{\mathcal{A}}_i^{k+1}(\theta) = \begin{cases} \tilde{\mathcal{L}}_i(\theta; \theta^{(k)}, \{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}), & \text{if } i = i_k \\ \tilde{\mathcal{A}}_i^k(\theta), & \text{otherwise.} \end{cases} \quad (7)$$

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

125 and applying the Jensen inequality:

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) = \int_{\mathcal{Z}} \underbrace{\log(p_i(z_i, \bar{\theta})/f_i(z_i, \theta))}_{=r_i(\theta; \bar{\theta}, z_i)} p_i(z_i, \bar{\theta}) \mu_i(dz_i), \quad (9)$$

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

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

$$p(y, w|x) = \pi(w) \prod_{i=1}^n p(y_i|x_i, w). \quad (10)$$

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

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) := \text{KL}(q(w; \theta) || p(w|y, x)) := \mathbb{E}_{q(w; \theta)} [\log(q(w; \theta)/p(w|y, x))] . \quad (11)$$

135 Using (10), we decompose  $\mathcal{L}(\theta) = n^{-1} \sum_{i=1}^n \mathcal{L}_i(\theta) + \text{const.}$  where:

$$\mathcal{L}_i(\theta) := -\mathbb{E}_{q(w; \theta)} [\log p(y_i|x_i, w)] + \frac{1}{n} \mathbb{E}_{q(w; \theta)} [\log q(w; \theta)/\pi(w)] = r_i(\theta) + d(\theta). \quad (12)$$

136 Directly optimizing the finite sum objective function in (11) can be difficult. First, with  $n \gg 1$ ,  
137 evaluating the objective function  $\mathcal{L}(\theta)$  requires a full pass over the entire dataset. Second, for some  
138 complex models, the expectations in (12) can be intractable even if we assume a simple parametric  
139 model for  $q(w; \theta)$ . Assume that  $\mathcal{L}_i$  is L-smooth, i.e.,  $\mathcal{L}_i$  is differentiable on  $\Theta$  and its gradient  $\nabla \mathcal{L}_i$   
140 is L-Lipschitz. We apply the MISSO method with a quadratic surrogate function defined as:

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) := \mathcal{L}_i(\bar{\theta}) + \langle \nabla_{\theta} \mathcal{L}_i(\bar{\theta}) | \theta - \bar{\theta} \rangle + \frac{L}{2} \|\bar{\theta} - \theta\|^2. \quad (13)$$

141 It is easily checked that  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  satisfies S1, S2.

142 To compute the gradient  $\nabla \mathcal{L}_i(\bar{\theta})$ , we apply the re-parametrization technique suggested in [Paisley  
143 et al., 2012, Kingma and Welling, 2014, Blundell et al., 2015]. Let  $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$  be a differen-  
144 tiable function w.r.t.  $\theta \in \Theta$  which is designed such that the law of  $w = t(z, \bar{\theta})$ , where  $z \sim \mathcal{N}_d(0, \mathbf{I})$ ,  
145 is  $q(\cdot, \bar{\theta})$ . By [Blundell et al., 2015, Proposition 1], the gradient of  $-r_i(\cdot)$  in (12) is:

$$\nabla_{\theta} \mathbb{E}_{q(w; \bar{\theta})} [\log p(y_i|x_i, w)] = \mathbb{E}_{z \sim \mathcal{N}_d(0, \mathbf{I})} [\mathbf{J}_{\theta}^t(z, \bar{\theta}) \nabla_w \log p(y_i|x_i, w)|_{w=t(z, \bar{\theta})}], \quad (14)$$

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

$$r_i(\theta; \bar{\theta}, z) := \left\langle \nabla_{\theta} d(\bar{\theta}) - J_{\theta}^t(z, \bar{\theta}) \nabla_w \log p(y_i | x_i, w) \Big|_{w=t(z, \bar{\theta})} \mid \theta - \bar{\theta} \right\rangle + \frac{L}{2} \|\theta - \bar{\theta}\|^2. \quad (15)$$

Finally, using (13) and (15), the surrogate function (5) is given by  $\tilde{\mathcal{L}}_i(\theta; \bar{\theta}, \{z_m\}_{m=1}^M) := M^{-1} \sum_{m=1}^M r_i(\theta; \bar{\theta}, z_m)$  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.

**H1.** For all  $i \in \llbracket 1, n \rrbracket$ ,  $\bar{\theta} \in \Theta$ ,  $z_i \in \mathcal{Z}$ , the measurable function  $r_i(\theta; \bar{\theta}, z_i)$  is convex in  $\theta$  and is lower bounded.

We are particularly interested in the *constrained optimization* setting where  $\Theta$  is a bounded set. To this end, we control the supremum norm of the of the above approximation as:

**H2.** For all  $i \in \llbracket 1, n \rrbracket$ ,  $(\theta, \bar{\theta}) \in \Theta^2$ ,  $z_i \in \mathcal{Z}$  we assume the existence of a majorizing function  $m_r : \mathcal{Z} \rightarrow \mathbb{R}$  and a constant  $C_r < \infty$  such that:

$$\sup_{M>0} \frac{1}{\sqrt{M}} \left| \sum_{m=1}^M \left\{ r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta}) \right\} \right| < m_r(z_i) \quad \text{and} \quad \mathbb{E}_{\bar{\theta}}[m_r(z_i) | \mathcal{F}] < C_r \quad (16)$$

where  $\mathcal{F}$  is the filtration of the total randomness and we denoted by  $\mathbb{E}_{\bar{\theta}}[\cdot]$  the expectation w.r.t. a Markov chain  $\{z_{i,m}\}_{m=1}^M$  with initial distribution  $\xi_i(\cdot; \bar{\theta})$ , transition kernel  $P_{i, \bar{\theta}}$ , and stationary distribution  $p_i(\cdot; \bar{\theta})$ . Besides, there exists a majorizing function  $m_{gr} : \mathcal{Z} \rightarrow \mathbb{R}$  and a constant  $C_{gr} < \infty$  such that:

$$\sup_{M>0} \frac{1}{\sqrt{M}} \left| \sum_{m=1}^M \left\{ \frac{\hat{\mathcal{L}}'_i(\theta, \theta - \bar{\theta}; \bar{\theta}) - r'_i(\theta, \theta - \bar{\theta}; \bar{\theta}, z_{i,m})}{\|\bar{\theta} - \theta\|} \right\} \right| < m_{gr}(z_i) \quad (17)$$

$$\mathbb{E}_{\bar{\theta}}[m_{gr}(z_i) | \mathcal{F}] < C_{gr}$$

**Some intuitions behind the controlling terms:** It is actually common in statistical and optimization problems, to deal with the manipulation and the control of random variables indexed by sets with an infinite number of elements. here, the random variable we control is an image of a continuous function noted  $v : \mathcal{Z} \rightarrow \mathbb{R}$  and defined as  $v(z) := r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta})$  for all  $z \in \mathcal{Z}$  and for fixed  $(\theta, \bar{\theta}) \in \Theta^2$ . To characterize such control, we will have recourse to the notion of metric entropy (or covering number of bracketing number) as developed in [Van der Vaart, 2000, Vershynin, 2018, Wainwright, 2019]. A collection of results from those books gives intuition behind our assumption H2, classical in empirical process:

In [Vershynin, 2018], the authors recall the uniform law of large numbers by stating that for  $(X_i, i \in \llbracket 1, M \rrbracket)$  random variables taking values in  $(0, 1)$ , we have:

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{M} \sum_{i=1}^M f(X_i) - \mathbb{E} f(X) \right| \leq \frac{CL}{\sqrt{M}} \quad (18)$$

Moreover, in [Vershynin, 2018] and [Wainwright, 2019], the application of the Dudley's inequality yields:

$$\mathbb{E} \sup_f |X_f| = \mathbb{E} \sup_f |X_f - X_0| \leq \frac{1}{\sqrt{M}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{\infty}, \varepsilon)} d\varepsilon \quad (19)$$

where  $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$  is the bracketing number and  $\varepsilon$  denotes the level of approximation (the bracketing number goes to infinity when  $\varepsilon \rightarrow 0$ ). Finally, in [Van der Vaart, 2000], this bracketing number is upperbounded for a class of parametric function  $\mathcal{F} = f_\theta : \theta \in \Theta$  on a bounded set  $\Theta \subset \mathbb{R}$  as:

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq K \left( \frac{\text{diam } \Theta}{\varepsilon} \right)^d, \quad \text{every } 0 < \varepsilon < \text{diam } \Theta \quad (20)$$

It is worth contrasting the exponential dependence of this metric entropy on the dimension  $d$ . The authors acknowledge that this is a dramatic manifestation of the curse of dimensionality happening when sampling is needed.

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

$$g(\bar{\theta}) := \inf_{\theta \in \Theta} \frac{\mathcal{L}'(\bar{\theta}, \theta - \bar{\theta})}{\|\bar{\theta} - \theta\|} \quad \text{and} \quad g(\bar{\theta}) = g_+(\bar{\theta}) - g_-(\bar{\theta}), \quad (21)$$

where  $g_+(\bar{\theta}) := \max\{0, g(\bar{\theta})\}$ ,  $g_-(\bar{\theta}) := -\min\{0, g(\bar{\theta})\}$  denote the positive and negative part of  $g(\bar{\theta})$ , respectively. Note that  $\bar{\theta}$  is a stationary point if and only if  $g_-(\bar{\theta}) = 0$  [Fletcher et al., 2002].

Also, denote

$$\widehat{\mathcal{L}}^{(k)}(\theta) := \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}_i(\theta; \theta^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\theta) := \widehat{\mathcal{L}}^{(k)}(\theta) - \mathcal{L}(\theta). \quad (22)$$

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

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

$$\Delta_{(K_{\max})} := 2n\mathbb{E}[\widehat{\mathcal{L}}^{(0)}(\theta^{(0)}) - \widehat{\mathcal{L}}^{(K_{\max})}(\theta^{(K_{\max})})] + \sum_{k=0}^{K_{\max}-1} \frac{4LC_r}{\sqrt{M_{(k)}}}, \quad (23)$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\theta^{(K)})\|^2] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \quad (24)$$

$$\mathbb{E}[g_-(\theta^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}. \quad (25)$$

Note that  $\Delta_{(K_{\max})}$  is finite for any  $K_{\max} \in \mathbb{N}$ . As expected, the MISSO method converges to a stationary point of (1) asymptotically and at a sublinear rate  $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\max}})$ .

Furthermore, we remark that the MISO method can be analyzed in Theorem 1 as a special case of the MISSO method satisfying  $C_r = C_{\text{gr}} = 0$ . In this case, while the asymptotic convergence is well known from [Mairal, 2015] [cf. H2], Eq. (24) gives a non-asymptotic rate of  $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{nL/K_{\max}})$  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  $\{M_{(k)}\}_{k \geq 0}$  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.,  $\lim_{k \rightarrow \infty} g_-(\theta^{(k)}) = 0$  a.s.
2. the objective value  $\mathcal{L}(\theta^{(k)})$  converges almost surely to a finite number  $\underline{\mathcal{L}}$ , i.e.,  $\lim_{k \rightarrow \infty} \mathcal{L}(\theta^{(k)}) = \underline{\mathcal{L}}$  a.s.

In particular, the first result above shows that the sequence  $\{\theta^{(k)}\}_{k \geq 0}$  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,  $((y_i, z_i), i \in \llbracket n \rrbracket)$  where  $y_i \in \{0, 1\}$  is a binary response and  $z_i = (z_{i,j} \in \mathbb{R}, j \in \llbracket p \rrbracket)$  is a covariate vector. The vector of covariates  $z_i = [z_{i,\text{mis}}, z_{i,\text{obs}}]$  is not fully observed where we denote by  $z_{i,\text{mis}}$  the missing values and  $z_{i,\text{obs}}$  the observed covariate. It is assumed that  $(z_i, i \in \llbracket n \rrbracket)$  are i.i.d. and marginally distributed according to  $\mathcal{N}(\beta, \Omega)$  where  $\beta \in \mathbb{R}^p$  and  $\Omega$  is a positive definite  $p \times p$  matrix.

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

$$p_i(y_i|z_i) = S(\delta^\top \bar{z}_i)^{y_i} (1 - S(\delta^\top \bar{z}_i))^{1-y_i} \quad (26)$$

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

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

**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. Details on the surrogate functions and the parameters updates are given in (83) and Appendix D.1.3.

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 (26) 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  $z_{i,\text{mis}}$ , we run a Metropolis Hastings algorithm with the target distribution  $p(\cdot|z_{i,\text{obs}}, y_i; \theta^{(k)})$  whose procedure is detailed in Appendix D.1.2.

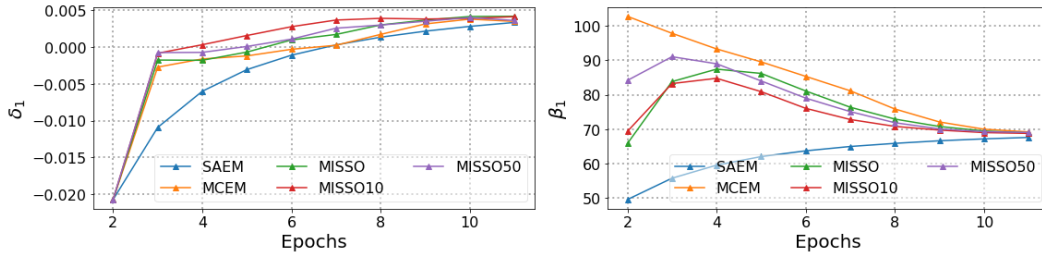


Figure 1: Convergence of first component of the vector of parameters  $\delta$  and  $\beta$  for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against the number of passes over the data.

We compare in Figure 1 the convergence behavior of the estimated parameters  $\beta$  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 Training Bayesian CNN using MISSO

At iteration  $k$ , minimizing the sum of stochastic surrogates defined as in (5) and (15) yields the following MISSO update — step (i) pick a function index  $i_k$  uniformly on  $\llbracket n \rrbracket$ ; 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

$$\mu_\ell^{(k)} = \frac{1}{n} \sum_{i=1}^n \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\mu_\ell, i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n} \sum_{i=1}^n \sigma^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\sigma, i}^{(k)}, \quad (27)$$

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

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

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

**Bayesian LeNet-5 on MNIST [LeCun et al., 1998]:** 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,  $28 \times 28$  images. Each image is labelled with its corresponding number (from zero to nine). Under the prior distribution  $\pi$ , see (10), 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  $\mu = (\mu_\ell, \ell \in \llbracket d \rrbracket)$  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  $z \sim \mathcal{N}(0, \mathbf{I})$ .

We describe in Table 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 ( $5 \times 5$ )	6	1	0	$1 \times 32 \times 32$	ReLU
max-pooling ( $2 \times 2$ )		2	0	$6 \times 28 \times 28$	
convolution ( $5 \times 5$ )	6	1	0	$1 \times 14 \times 14$	ReLU
max-pooling ( $2 \times 2$ )		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

**Bayesian ResNet-18 [He et al., 2016] on CIFAR-10 [Krizhevsky et al., 2012]:** We train here the Bayesian variant of the ResNet-18 neural network (see Appendix D.2.2) introduced in [He et al., 2016] on CIFAR-10. The latter dataset is composed of  $n = 60\,000$  handwritten digits,  $32 \times 32$  colour images in 10 classes, with 6000 images per class. As in the previous example, the weights are assumed independent and identically distributed according to  $\mathcal{N}(0, 1)$ . The source code used as a backbone here can be found in the TensorFlow Probability Github repo ([https://github.com/tensorflow/probability/blob/master/tensorflow\\_probability/examples/cifar10\\_bnn.py](https://github.com/tensorflow/probability/blob/master/tensorflow_probability/examples/cifar10_bnn.py)) where the default hyperparameters, as the L annealing constant or the number of MC samples, were used for the benchmark methods. For better efficiency and lower variance, the Flipout estimator [Wen et al., 2018] is preferred than a simple reparametrization trick for ResNet-18.

We describe in Table 2 the architecture of the Resnet-18 we train on CIFAR-10:



layer type	Output Size	ResNet-18	nonlinearity
conv1	$112 \times 112 \times 64$	$7 \times 7, 64, \text{stride } 2$	ReLU
conv2x	$56 \times 56 \times 64$	$\begin{pmatrix} 3 \times 3, 64 \\ 3 \times 3, 64 \end{pmatrix} \times 2$	ReLU
conv3x	$28 \times 28 \times 128$	$\begin{pmatrix} 3 \times 3, 128 \\ 3 \times 3, 128 \end{pmatrix} \times 2$	ReLU
conv4x	$14 \times 14 \times 256$	$\begin{pmatrix} 3 \times 3, 256 \\ 3 \times 3, 256 \end{pmatrix} \times 2$	ReLU
conv5x	$7 \times 7 \times 512$	$\begin{pmatrix} 3 \times 3, 512 \\ 3 \times 3, 512 \end{pmatrix} \times 2$	ReLU
average pool	$1 \times 1 \times 512$	$7 \times 7$ average pool	ReLU
fully connected	1000	$512 \times 1000$ fully connections	
softmax	1000		

Table 2: ResNet-18 architecture

**Experiment Results:** 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 (12) 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.2.3. 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  $M_{(k)} = k$ .

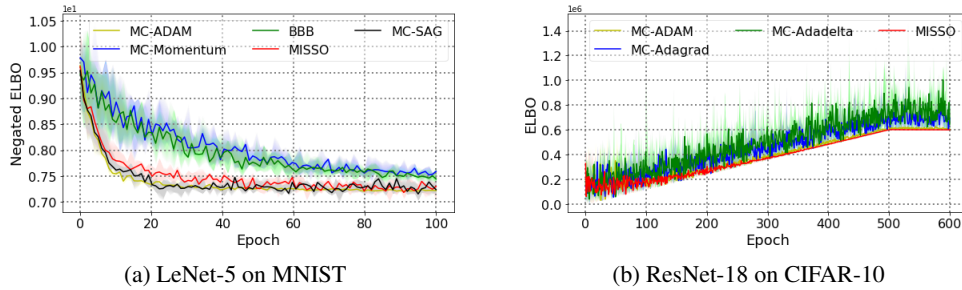


Figure 2: (a) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. (b) ELBO versus epochs elapsed for fitting the Bayesian ResNet-18 on CIFAR-10 using different algorithms. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

Figure 2(a) 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 for our experiment on MNIST dataset using a Bayesian variant of LeNet-5. On the other hand, the experiment conducted on CIFAR-10 (Figure 2(b)) using a much larger network, *i.e.*, a Bayesian variant of ResNet-18 showcases the need of a well-tuned adaptive methods to reach better training loss (and also faster). Our MISSO method is similar to the Monte Carlo variant of ADAM but slower than built-in TF optimizers such as Adadelta and Adagrad. Recall that the purpose of this paper is to provide a common class of optimizers, such as VI, in order to study their convergence behaviors and not to introduce a novel method outperforming the rest.

## 285 **5 Conclusion**

286 We present a unifying framework for minimizing a non-convex finite-sum objective function using  
287 incremental surrogates when the latter functions are expressed as an expectation and are intractable.  
288 Our approach covers a large class of non-convex applications in machine learning such as logistic  
289 regression with missing values and variational inference. We provide both finite-time and asymptotic  
290 guarantees of our incremental stochastic surrogate optimization technique and illustrate our findings  
291 training a binary logistic regression with missing covariates to predict hemorrhagic shock and a  
292 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.
- 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>.
- 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.
- K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- 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>.
- 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>.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- 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.

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

342 J. Mairal. Incremental majorization-minimization optimization with application to large-scale ma-  
343 chine learning. *SIAM J. Optim.*, 25(2):829–855, 2015. ISSN 1052-6234. doi: 10.1137/  
344 140957639. URL <https://doi.org/10.1137/140957639>.

345 G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley Series in Probabil-  
346 ity and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.  
347 ISBN 978-0-471-20170-0. doi: 10.1002/9780470191613. URL [https://doi.org/10.1002/](https://doi.org/10.1002/9780470191613)  
348 [9780470191613](https://doi.org/10.1002/9780470191613).

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

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

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

355 N. G. Polson, V. Sokolov, et al. Deep learning: a bayesian perspective. *Bayesian Analysis*, 12(4):  
356 1275–1304, 2017.

357 X. Qian, A. Sailanbayev, K. Mishchenko, and P. Richtárik. Miso is making a comeback with better  
358 proofs and rates. *arXiv preprint arXiv:1906.01474*, 2019.

359 M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive  
360 minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–  
361 1153, 2013.

362 D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate in-  
363 ference in deep generative models. In *International Conference on Machine Learning*, pages  
364 1278–1286, 2014.

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

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

369 A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

370 R. Vershynin. *High-dimensional probability: An introduction with applications in data science*,  
371 volume 47. Cambridge university press, 2018.

372 M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge  
373 University Press, 2019.

374 G. C. G. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor  
375 man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):  
376 699–704, 1990. doi: 10.1080/01621459.1990.10474930. URL [https://www.tandfonline.](https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474930)  
377 [com/doi/abs/10.1080/01621459.1990.10474930](https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474930).

378 Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse. Flipout: Efficient pseudo-independent weight  
379 perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

## 380 A Proof of Theorem 1

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

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

383 Then we have following non-asymptotic bounds:

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

384 **Proof** We begin by recalling the definition

$$\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{A}}_i^k(\boldsymbol{\theta}). \quad (28)$$

385 Notice that

$$\begin{aligned} \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^{k+1})}, \{z_{i,m}^{(\tau_i^{k+1})}\}_{m=1}^{M_{(\tau_i^{k+1})}}) \\ &= \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n} (\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})). \end{aligned} \quad (29)$$

386 Furthermore, we recall that

$$\hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \hat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \hat{e}^{(k)}(\boldsymbol{\theta}) := \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \quad (30)$$

387 Due to S2, we have

$$\|\nabla \hat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\hat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \quad (31)$$

388 To prove the first bound in (24), using the optimality of  $\boldsymbol{\theta}^{(k+1)}$ , one has

$$\begin{aligned} \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) &\leq \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\ &= \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{n} (\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})) \end{aligned} \quad (32)$$

389 Let  $\mathcal{F}_k$  be the filtration of random variables up to iteration  $k$ , i.e.,  $\{i_{\ell-1}, \{z_{i_{\ell-1},m}^{(\ell-1)}\}_{m=1}^{M_{(\ell-1)}}, \boldsymbol{\theta}^{(\ell)}\}_{\ell=1}^k$ .

390 We observe that the conditional expectation evaluates to

$$\begin{aligned} \mathbb{E}_{i_k} [\mathbb{E}[\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) | \mathcal{F}_k, i_k] | \mathcal{F}_k] \\ = \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathbb{E}_{i_k} [\mathbb{E}[\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} r_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, z_{i_k,m}^{(k)}) - \hat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) | \mathcal{F}_k, i_k] | \mathcal{F}_k] \\ \leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{C_r}{\sqrt{M_{(k)}}}, \end{aligned} \quad (33)$$

391 where the last inequality is due to H2. Moreover,

$$\mathbb{E}[\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}}) | \mathcal{F}_k] = \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, \{z_{i,m}^{(\tau_i^k)}\}_{m=1}^{M_{(\tau_i^k)}}) = \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}). \quad (34)$$

392 Taking the conditional expectations on both sides of (32) and re-arranging terms give:

$$\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) | \mathcal{F}_k] + \frac{C_r}{\sqrt{M_{(k)}}} \quad (35)$$

Proceeding from (35), we observe the following lower bound for the left hand side

$$\begin{aligned}
& \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \stackrel{(a)}{=} \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) \\
& \stackrel{(b)}{\geq} \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2L} \|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
& = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) - \hat{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \right\} + \frac{1}{2L} \|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
& \quad \underbrace{\hspace{10em}}_{:= -\delta^{(k)}(\boldsymbol{\theta}^{(k)})}
\end{aligned} \tag{36}$$

where (a) is due to  $\hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$  [cf. S1], (b) is due to (31) and we have defined the summation in the last equality as  $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$ . Substituting the above into (35) yields

$$\frac{\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \leq n \mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) | \mathcal{F}_k] + \frac{C_r}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)}) \tag{37}$$

Observe the following upper bound on the total expectations:

$$\mathbb{E}[\delta^{(k)}(\boldsymbol{\theta}^{(k)})] \leq \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \frac{C_r}{\sqrt{M_{(\tau_i^k)}}}\right], \tag{38}$$

which is due to H2. It yields

$$\mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \leq 2nL \mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})] + \frac{2LC_r}{\sqrt{M_{(k)}}} + \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\frac{2LC_r}{\sqrt{M_{(\tau_i^k)}}}\right]$$

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

$$\begin{aligned}
\mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] &= \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL \mathbb{E}[\tilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \tilde{\mathcal{L}}^{(K_{\max})}(\boldsymbol{\theta}^{(K_{\max})})]}{K_{\max}} + \frac{2LC_r}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{M_{(\tau_i^k)}}}\right]
\end{aligned} \tag{39}$$

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:

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^k \frac{1}{n} \left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} \tag{40}$$

Taking the sum yields:

$$\begin{aligned}
\sum_{k=0}^{K_{\max}-1} \mathbb{E}[M_{(\tau_i^k)}^{-1/2}] &= \sum_{k=0}^{K_{\max}-1} \sum_{j=1}^k \frac{1}{n} \left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\max}-1} \sum_{l=0}^{k-1} \frac{1}{n} \left(1 - \frac{1}{n}\right)^{k-(l+1)} M_{(l)}^{-1/2} \\
&= \sum_{l=0}^{K_{\max}-1} M_{(l)}^{-1/2} \sum_{k=l+1}^{K_{\max}-1} \frac{1}{n} \left(1 - \frac{1}{n}\right)^{k-(l+1)} \leq \sum_{l=0}^{K_{\max}-1} M_{(l)}^{-1/2}
\end{aligned} \tag{41}$$

where the last inequality is due to upper bounding the geometric series. Plugging this back into (39) yields

$$\begin{aligned}
\mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] &= \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL \mathbb{E}[\tilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \tilde{\mathcal{L}}^{(K_{\max})}(\boldsymbol{\theta}^{(K_{\max})})]}{K_{\max}} + \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \frac{4LC_r}{\sqrt{M_{(k)}}} = \frac{\Delta_{(K_{\max})}}{K_{\max}}.
\end{aligned} \tag{42}$$



405 This concludes our proof for the first inequality in (24).

406 To prove the second inequality of (24), we define the shorthand notations  $g^{(k)} := g(\theta^{(k)})$ ,  $g_-^{(k)} :=$   
 407  $-\min\{0, g^{(k)}\}$ ,  $g_+^{(k)} := \max\{0, g^{(k)}\}$ . We observe that

$$\begin{aligned} g^{(k)} &= \inf_{\theta \in \Theta} \frac{\mathcal{L}'(\theta^{(k)}, \theta - \theta^{(k)})}{\|\theta^{(k)} - \theta\|} \\ &= \inf_{\theta \in \Theta} \left\{ \frac{\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\theta^{(k)}, \theta - \theta^{(k)}; \theta^{(\tau_i^k)})}{\|\theta^{(k)} - \theta\|} - \frac{\langle \nabla \widehat{e}^{(k)}(\theta^{(k)}) | \theta - \theta^{(k)} \rangle}{\|\theta^{(k)} - \theta\|} \right\} \\ &\geq -\|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| + \inf_{\theta \in \Theta} \frac{\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\theta^{(k)}, \theta - \theta^{(k)}; \theta^{(\tau_i^k)})}{\|\theta^{(k)} - \theta\|} \end{aligned} \quad (43)$$

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

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\theta^{(k)}, \theta - \theta^{(k)}; \theta^{(\tau_i^k)}) \\ &= \underbrace{\widetilde{\mathcal{L}}^{(k)'}(\theta^{(k)}, \theta - \theta^{(k)}) - \widetilde{\mathcal{L}}^{(k)'}(\theta^{(k)}, \theta - \theta^{(k)})}_{\geq 0} + \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\theta^{(k)}, \theta - \theta^{(k)}; \theta^{(\tau_i^k)}) \\ &\geq \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{\mathcal{L}}'_i(\theta^{(k)}, \theta - \theta^{(k)}; \theta^{(\tau_i^k)}) - \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r'_i(\theta^{(k)}, \theta - \theta^{(k)}; \theta^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) \right\} \end{aligned} \quad (44)$$

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

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

413 We have

$$g^{(k)} \geq -\|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| + \inf_{\theta \in \Theta} (-\epsilon^{(k)}(\theta)) \geq -\|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| - \sup_{\theta \in \Theta} |\epsilon^{(k)}(\theta)|. \quad (45)$$

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

$$g_-^{(k)} \leq \|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| + \sup_{\theta \in \Theta} |\epsilon^{(k)}(\theta)|. \quad (46)$$

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

$$\mathbb{E}[g_-^{(K)}] \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\theta^{(K)})\|] + \mathbb{E}[\sup_{\theta \in \Theta} \epsilon^{(K)}(\theta)] \quad (47)$$

417 We note that

$$\left( \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\theta^{(K)})\|] \right)^2 \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\theta^{(K)})\|^2] \leq \frac{\Delta(K_{\max})}{K_{\max}}, \quad (48)$$

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

$$\begin{aligned} \mathbb{E}[\sup_{\theta \in \Theta} \epsilon^{(K)}(\theta)] &= \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}} \mathbb{E}[\sup_{\theta \in \Theta} \epsilon^{(k)}(\theta)] \stackrel{(a)}{\leq} \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n M_{(\tau_i^k)}^{-1/2}\right] \\ &\stackrel{(b)}{\leq} \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2} \end{aligned} \quad (49)$$

419 where (a) is due to H2 and (b) is due to (41). This implies

$$\mathbb{E}[g_-^{(K)}] \leq \sqrt{\frac{\Delta(K_{\max})}{K_{\max}}} + \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}, \quad (50)$$

420 and concludes the proof of the theorem.  $\square$

## 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.,  $\lim_{k \rightarrow \infty} g_{-}(\boldsymbol{\theta}^{(k)}) = 0$  a.s..
2. the objective value  $\mathcal{L}(\boldsymbol{\theta}^{(k)})$  converges almost surely to a finite number  $\underline{\mathcal{L}}$ , i.e.,  $\lim_{k \rightarrow \infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$  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  $k \geq 1$ :

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1} \quad (51)$$

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  $V_{\infty}$ .

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

(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 (32) by re-arranging terms and observing that

$$\begin{aligned} \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) &\leq \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad - (\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k, m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})) \end{aligned} \quad (52)$$

Our idea is to apply Lemma 1. Under S1, the finite sum of surrogate functions  $\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta})$ , defined in (22), is lower bounded by a constant  $c_k > -\infty$  for any  $\boldsymbol{\theta}$ . To this end, we observe that

$$V_k := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \inf_{k \geq 0} c_k \geq 0 \quad (53)$$

is a non-negative random variable.

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

$$X_k := \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(\tau_{i_k}^k)}; \boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \geq 0. \quad (54)$$

Thirdly, we define

$$\begin{aligned} E_k &= -(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k, m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})). \end{aligned} \quad (55)$$

Note that from the definitions (53), (54), (55), we have  $V_{k+1} \leq V_k - X_k + E_k$  for any  $k \geq 1$ .

Under H2, we observe that

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

$$\mathbb{E} \left[ \left| \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}}) \right| \right] \leq C_r \mathbb{E} \left[ M_{(\tau_{i_k}^k)}^{-1/2} \right] \quad (57)$$

$$\mathbb{E} \left[ \left| \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) \right| \right] \leq \frac{1}{n} \sum_{i=1}^n C_r \mathbb{E} \left[ M_{(\tau_i^k)}^{-1/2} \right] \quad (58)$$

Therefore,

$$\mathbb{E} [|E_k|] \leq \frac{C_r}{n} \left( M_{(k)}^{-1/2} + \mathbb{E} \left[ M_{(\tau_{i_k}^k)}^{-1/2} + \sum_{i=1}^n \{ M_{(\tau_i^k)}^{-1/2} + M_{(\tau_{i+1}^k)}^{-1/2} \} \right] \right) \quad (59)$$

Using (41) and the assumption on the sequence  $\{M_{(k)}\}_{k \geq 0}$ , we obtain that

$$\sum_{k=0}^{\infty} \mathbb{E} [|E_k|] < \frac{C_r}{n} (2 + 2n) \sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty. \quad (60)$$

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

$$\begin{aligned} \infty &> \sum_{k=0}^{\infty} \mathbb{E} [X_k] = \frac{1}{n} \sum_{k=0}^{\infty} \mathbb{E} [\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})] \\ &= \frac{1}{n} \sum_{k=0}^{\infty} \mathbb{E} [\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})] = \frac{1}{n} \sum_{k=0}^{\infty} \mathbb{E} [\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})] \end{aligned} \quad (61)$$

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

$$\lim_{k \rightarrow \infty} \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0 \quad \text{a.s.} \quad (62)$$

and subsequently applying (31), we have  $\lim_{k \rightarrow \infty} \|\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| = 0$  almost surely. Finally, it follows from (31) and (46) that

$$\lim_{k \rightarrow \infty} g_-^{(k)} \leq \lim_{k \rightarrow \infty} \sqrt{2L} \sqrt{\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})} + \lim_{k \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})| = 0, \quad (63)$$

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  $\lim_{k \rightarrow \infty} \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$ . Applying (62) implies that

$$\underline{\mathcal{L}} = \lim_{k \rightarrow \infty} \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \lim_{k \rightarrow \infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \quad \text{a.s.} \quad (64)$$

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

## 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  $V_{\infty}$ .
- (ii) the sequence  $(\mathbb{E}[V_k])_{k \geq 0}$  converges and  $\lim_{k \rightarrow \infty} \mathbb{E}[V_k] = \mathbb{E}[V_{\infty}]$ .
- (iii) the series  $\sum_{k=0}^{\infty} X_k$  converges almost surely and  $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$ .

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

$$0 \leq V_k \leq V_0 - \sum_{j=1}^k X_j + \sum_{j=1}^k E_j \leq V_0 + \sum_{j=1}^k E_j \quad (65)$$

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

470 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  
471  $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$ , the series  $\sum_{j=1}^{\infty} E_j$  converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \quad (66)$$

472 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:

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \end{aligned} \quad (67)$$

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

$$\mathbb{E}[\liminf_{k \rightarrow \infty} |W_k|] = \mathbb{E}[|W_{\infty}|] \leq \liminf_{k \rightarrow \infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \quad (68)$$

477 showing that the random variable  $W_{\infty}$  is integrable.

478 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  
479  $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$  and by the monotone convergence theorem, we get:

$$\lim_{k \rightarrow \infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k \rightarrow \infty} U_k] \quad (69)$$

480 Finally, we have:

$$\lim_{k \rightarrow \infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_{\infty} \quad \text{and} \quad \mathbb{E}[\lim_{k \rightarrow \infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_{\infty}] \quad (70)$$

481 showing that  $\mathbb{E}[W_{\infty}] = w_{\infty}$  and concluding the proof of (ii). Moreover, using (67) we have that  
482  $W_k \leq W_{k-1} - X_k$  which yields:

$$\begin{aligned} \sum_{j=1}^{\infty} X_j &\leq W_0 - W_{\infty} < \infty \\ \sum_{j=1}^{\infty} \mathbb{E}[X_j] &\leq \mathbb{E}[W_0] - w_{\infty} < \infty \end{aligned} \quad (71)$$

483 which concludes the proof of the lemma.  $\square$

## 484 D Details about the Numerical Experiments

### 485 D.1 Binary Logistic Regression on the Traumabase

#### 486 D.1.1 Traumabase quantitative variables

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

#### 495 D.1.2 Metropolis Hastings algorithm

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

$$\begin{aligned} \xi &= \beta_{\text{mis}} + \Omega_{\text{mis},\text{obs}} \Omega_{\text{obs},\text{obs}}^{-1} (z_{i,\text{obs}} - \beta_{\text{obs}}) , \\ \Sigma &= \Omega_{\text{mis},\text{mis}} + \Omega_{\text{mis},\text{obs}} \Omega_{\text{obs},\text{obs}}^{-1} \Omega_{\text{obs},\text{mis}} \end{aligned} \quad (72)$$

501 where we have used the Schur Complement of  $\Omega_{\text{obs},\text{obs}}$  in  $\Omega$  and noted  $\beta_{\text{mis}}$  (resp.  $\beta_{\text{obs}}$ ) the missing  
 502 (resp. observed) elements of  $\beta$ . The MH algorithm is summarized in Algorithm 3.

---

#### Algorithm 3 MH algorithm

---

```

1: Input: initialization  $z_{i,\text{mis},0} \sim q(z_{i,\text{mis}}; \delta)$ 
2: for  $m = 1, \dots, M$  do
3:   Sample  $z_{i,\text{mis},m} \sim q(z_{i,\text{mis}}; \delta)$ 
4:   Sample  $u \sim \mathcal{U}([0, 1])$ 
5:   Calculate the ratio  $r = \frac{\pi(z_{i,\text{mis},m}; \theta) / q(z_{i,\text{mis},m}; \delta)}{\pi(z_{i,\text{mis},m-1}; \theta) / q(z_{i,\text{mis},m-1}; \delta)}$ 
6:   if  $u < r$  then
7:     Accept  $z_{i,\text{mis},m}$ 
8:   else
9:      $z_{i,\text{mis},m} \leftarrow z_{i,\text{mis},m-1}$ 
10:  end if
11: end for
12: Output:  $z_{i,\text{mis},M}$ 

```

---

#### 503 D.1.3 MISSO Update

504 **Choice of surrogate function for MISO:** We recall the MISO deterministic surrogate defined in  
 505 (9):

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) = \int_{\mathcal{Z}} \log(p_i(z_{i,\text{mis}}, \bar{\theta}) / f_i(z_{i,\text{mis}}, \theta)) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_i) . \quad (73)$$

506 where  $\theta = (\delta, \beta, \Omega)$  and  $\bar{\theta} = (\bar{\delta}, \bar{\beta}, \bar{\Omega})$ . We adapt it to our missing covariates problem and decom-  
 507 pose the the surrogate function defined above into an observed and a missing part.

508 **Surrogate function decomposition** We adapt it to our missing covariates problem and decompose  
 509 the term depending on  $\theta$ , while  $\bar{\theta}$  is fixed, in two following parts leading to

$$\begin{aligned}
 \hat{\mathcal{L}}_i(\theta; \bar{\theta}) &= - \int_{\mathbf{Z}} \log f_i(z_{i,\text{mis}}, z_{i,\text{obs}}, \theta) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) \\
 &= - \int_{\mathbf{Z}} \log [p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) p_i(z_{i,\text{mis}}, \beta, \Omega)] p_i(z_i, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) \\
 &= - \underbrace{\int_{\mathbf{Z}} \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) p_i(z_i, \bar{\theta}) \mu_i(dz_{i,\text{mis}})}_{=\hat{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta})} - \underbrace{\int_{\mathbf{Z}} \log p_i(z_{i,\text{mis}}, \beta, \Omega) p_i(z_i, \bar{\theta}) \mu_i(dz_{i,\text{mis}})}_{=\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})}
 \end{aligned} \tag{74}$$

510 The mean  $\beta$  and the covariance  $\Omega$  of the latent structure can be estimated minimizing the sum of  
 511 MISSO surrogates  $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta}, \{z_m\}_{m=1}^M)$ , defined as MC approximation of  $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})$ , for all  
 512  $i \in \llbracket n \rrbracket$ , in closed-form expression.

513 We thus keep the surrogate  $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})$  as it is, and consider the following quadratic approximation  
 514 of  $\hat{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta})$  to estimate the vector of logistic parameters  $\delta$ :

$$\begin{aligned}
 \hat{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta}) &- \int_{\mathbf{Z}} \nabla \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) \big|_{\delta=\bar{\delta}} p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) (\delta - \bar{\delta}) \\
 &- (\delta - \bar{\delta})/2 \int_{\mathbf{Z}} \nabla^2 \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) p_i(z_{i,\text{mis}}, \bar{\theta}) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) (\delta - \bar{\delta})^\top
 \end{aligned} \tag{75}$$

515 Recall that:

$$\begin{aligned}
 \nabla \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) &= z_i (y_i - S(\delta^\top z_i)) \\
 \nabla^2 \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) &= -z_i z_i^\top \dot{S}(\delta^\top z_i)
 \end{aligned} \tag{76}$$

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

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

520 Then, we use, for all  $i \in \llbracket n \rrbracket$ , the following surrogate function to estimate  $\delta$ :

$$\bar{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta}) = \hat{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta}) - D_i^\top (\delta - \bar{\delta}) + \frac{1}{2} (\delta - \bar{\delta}) H_i (\delta - \bar{\delta})^\top \tag{77}$$

521 where:

$$\begin{aligned}
 D_i &= \int_{\mathbf{Z}} \nabla \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) \big|_{\delta=\bar{\delta}} p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) \\
 H_i &= \int_{\mathbf{Z}} H_i(z_{i,\text{mis}}) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}})
 \end{aligned} \tag{78}$$

522 Finally, at iteration  $k$ , the total surrogate is:

$$\begin{aligned}
 \tilde{\mathcal{L}}^{(k)}(\theta) &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i(\theta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) \\
 &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) - \frac{1}{n} \sum_{i=1}^n \tilde{D}_i^{(\tau_i^k)} (\delta - \delta^{(\tau_i^k)}) \\
 &\quad + \frac{1}{2n} \sum_{i=1}^n (\delta - \delta^{(\tau_i^k)}) \left\{ \tilde{H}_i^{(\tau_i^k)} \right\} (\delta - \delta^{(\tau_i^k)})^\top
 \end{aligned} \tag{79}$$



523 where for all  $i \in \llbracket n \rrbracket$ :

$$\begin{aligned}\tilde{D}_i^{(\tau_i^k)} &= \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(\tau_i^k)} \left( y_i - S((\delta^{(\tau_i^k)})^\top z_{i,m}^{(\tau_i^k)}) \right) \\ \tilde{H}_i^{(\tau_i^k)} &= \frac{1}{4M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(\tau_i^k)} (z_{i,m}^{(\tau_i^k)})^\top\end{aligned}\tag{80}$$

524 Minimizing the total surrogate (79) boils down to performing a quasi-Newton step. It is perhaps sen-  
525 sible to apply some diagonal loading which is perfectly compatible with the surrogate interpretation  
526 we just gave.

527 The logistic parameters are estimated as follows:

$$\delta^{(k)} = \arg \min_{\delta \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(1)}(\delta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}})\tag{81}$$

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

$$\delta^{(k)} = \frac{1}{n} \sum_{i=1}^n \delta^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1} \tilde{D}^{(k)}\tag{82}$$

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

531 **MISSO updates:** At the  $k$ -th iteration, and after the initialization, for all  $i \in \llbracket n \rrbracket$ , of the latent  
532 variables  $(z_i^{(0)})$ , the MISSO algorithm consists in picking an index  $i_k$  uniformly on  $\llbracket n \rrbracket$ , complet-  
533 ing the observations by sampling a Monte Carlo batch  $\{z_{i_k, \text{mis}, m}^{(k)}\}_{m=1}^{M_{(k)}}$  of missing values from the  
534 conditional distribution  $p(z_{i_k, \text{mis}} | z_{i_k, \text{obs}}, y_{i_k}; \theta^{(k-1)})$  using an MCMC sampler and computing the  
535 estimated parameters as follows:

$$\begin{aligned}\beta^{(k)} &= \arg \min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega^{(k)}, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(k)} \\ \Omega^{(k)} &= \arg \min_{\Omega \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta^{(k)}, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} w_{i,m}^{(k)} \\ \delta^{(k)} &= \frac{1}{n} \sum_{i=1}^n \delta^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1} \tilde{D}^{(k)}.\end{aligned}\tag{83}$$

536 where  $z_{i,m}^{(k)} = (z_{i, \text{mis}, m}^{(k)}, z_{i, \text{obs}})$  is composed of a simulated and an observed part,  $\tilde{D}^{(k)} =$   
537  $\frac{1}{n} \sum_{i=1}^n \tilde{D}_i^{(\tau_i^k)}$ ,  $\tilde{H}^{(k)} = \frac{1}{n} \sum_{i=1}^n \tilde{H}_i^{(\tau_i^k)}$  and  $w_{i,m}^{(k)} = z_{i,m}^{(k)} (z_{i,m}^{(k)})^\top - \beta^{(k)} (\beta^{(k)})^\top$ . Be-  
538 sides,  $\tilde{\mathcal{L}}_i^{(1)}(\beta, \Omega, \bar{\theta}, \{z_m\}_{m=1}^M)$  and  $\tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta}, \{z_m\}_{m=1}^M)$  are defined as MC approximation of  
539  $\hat{\mathcal{L}}_i^{(1)}(\beta, \Omega, \bar{\theta})$  and  $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})$ , for all  $i \in \llbracket n \rrbracket$  as components of the surrogate function (74).

## 540 D.2 Incremental Variational Inference

### 541 D.2.1 Bayesian LeNet-5 Architecture

542 [put here the table of the architecture](#)

### 543 D.2.2 Bayesian ResNet-18 Architecture

544 [put here the table of the architecture](#)

### 545 D.2.3 Algorithms updates

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

$$\begin{aligned}\hat{\delta}_{\mu_\ell, i}^{(k)} &= -\frac{1}{M^{(k)}} \sum_{m=1}^{M^{(k)}} \nabla_w \log p(y_i | x_i, w) \Big|_{w=t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} d(\boldsymbol{\theta}^{(k-1)}) , \\ \hat{\delta}_{\sigma, i}^{(k)} &= -\frac{1}{M^{(k)}} \sum_{m=1}^{M^{(k)}} z_m^{(k)} \nabla_w \log p(y_i | x_i, w) \Big|_{w=t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_\sigma d(\boldsymbol{\theta}^{(k-1)}) .\end{aligned}\tag{84}$$

548 For all benchmark algorithms, we pick, at iteration  $k$ , a function index  $i_k$  uniformly on  $\llbracket n \rrbracket$  and  
 549 sample a Monte Carlo batch  $\{z_m^{(k)}\}_{m=1}^{M^{(k)}}$  from the standard Gaussian distribution. The updates of the  
 550 parameters  $\mu_\ell$  for all  $\ell \in \llbracket d \rrbracket$  and  $\sigma$  break down as follows:

551 **Monte Carlo SAG update:** Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\mu_\ell, i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\sigma, i}^{(k)} ,\tag{85}$$

552 where  $\hat{\delta}_{\mu_\ell, i}^{(k)} = \hat{\delta}_{\mu_\ell, i}^{(k-1)}$  and  $\hat{\delta}_{\sigma, i}^{(k)} = \hat{\delta}_{\sigma, i}^{(k-1)}$  for  $i \neq i_k$  and are defined by (84) for  $i = i_k$ . The learning  
 553 rate is set to  $\gamma = 10^{-3}$ .

554 **Bayes By Backprop update:** Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n} \hat{\delta}_{\mu_\ell, i_k}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n} \hat{\delta}_{\sigma, i_k}^{(k)} ,\tag{86}$$

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

556 **Monte Carlo Momentum update:** Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} + \hat{\mathbf{v}}_{\mu_\ell}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} + \hat{\mathbf{v}}_\sigma^{(k)} ,\tag{87}$$

557 where

$$\hat{\mathbf{v}}_{\mu_\ell, i}^{(k)} = \alpha \hat{\mathbf{v}}_{\mu_\ell, i}^{(k-1)} - \frac{\gamma}{n} \hat{\delta}_{\mu_\ell, i_k}^{(k)} \quad \text{and} \quad \hat{\mathbf{v}}_\sigma^{(k)} = \alpha \hat{\mathbf{v}}_\sigma^{(k-1)} - \frac{\gamma}{n} \hat{\delta}_{\sigma, i_k}^{(k)} ,\tag{88}$$

558 where  $\alpha$  and  $\gamma$ , respectively the momentum and the learning rates, are set to  $10^{-3}$ .

559 **Monte Carlo ADAM update:** Set

$$\mu_\ell^{(k)} = \mu_\ell^{(k-1)} - \frac{\gamma}{n} \hat{\mathbf{m}}_{\mu_\ell}^{(k)} / (\sqrt{\hat{\mathbf{m}}_{\mu_\ell}^{(k)}} + \epsilon) \quad \text{and} \quad \sigma^{(k)} = \sigma^{(k-1)} - \frac{\gamma}{n} \hat{\mathbf{m}}_\sigma^{(k)} / (\sqrt{\hat{\mathbf{m}}_\sigma^{(k)}} + \epsilon) ,\tag{89}$$

560 where

$$\begin{aligned}\hat{\mathbf{m}}_{\mu_\ell}^{(k)} &= \mathbf{m}_{\mu_\ell}^{(k-1)} / (1 - \rho_1^k) \quad \text{with} \quad \mathbf{m}_{\mu_\ell}^{(k)} = \rho_1 \mathbf{m}_{\mu_\ell}^{(k-1)} + (1 - \rho_1) \hat{\delta}_{\mu_\ell, i_k}^{(k)} , \\ \hat{\mathbf{v}}_{\mu_\ell}^{(k)} &= \mathbf{v}_{\mu_\ell}^{(k-1)} / (1 - \rho_2^k) \quad \text{with} \quad \mathbf{v}_{\mu_\ell}^{(k)} = \rho_2 \mathbf{v}_{\mu_\ell}^{(k-1)} + (1 - \rho_2) (\hat{\delta}_{\mu_\ell, i_k}^{(k)})^2\end{aligned}\tag{90}$$

561 and

$$\begin{aligned}\hat{\mathbf{m}}_\sigma^{(k)} &= \mathbf{m}_\sigma^{(k-1)} / (1 - \rho_1^k) \quad \text{with} \quad \mathbf{m}_\sigma^{(k)} = \rho_1 \mathbf{m}_\sigma^{(k-1)} + (1 - \rho_1) \hat{\delta}_{\sigma, i_k}^{(k)} , \\ \hat{\mathbf{v}}_\sigma^{(k)} &= \mathbf{v}_\sigma^{(k-1)} / (1 - \rho_2^k) \quad \text{with} \quad \mathbf{v}_\sigma^{(k)} = \rho_2 \mathbf{v}_\sigma^{(k-1)} + (1 - \rho_2) (\hat{\delta}_{\sigma, i_k}^{(k)})^2 .\end{aligned}\tag{91}$$

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