

---

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

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

Many constrained, nonconvex and nonsmooth 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, nonconvex, nonsmooth optimization setting. We apply our new framework for inference of logistic regression model with missing data and for variational inference of Bayesian variants of LeNet-5 and Resnet-18 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) nonconvex and nonsmooth.

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 has 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]. A notable application of MISO-like algorithms is described in [?] where the authors build upon the stochastic majorization-minimization framework of

34 **Mairal [2015]** to introduce a method for sparse matrix factorization. Yet, in many applications of in-  
 35 terest, the natural surrogate functions are intractable, yet they are defined as expectation of tractable  
 36 functions. For instance, this is the case for inference in latent variable models via maximum likeli-  
 37 hood [**McLachlan and Krishnan, 2008**]. Another application is variational inference [**Ghahramani,**  
 38 **2015**], in which the goal is to approximate the posterior distribution of parameters given the obser-  
 39 vations; see for example [**Neal, 2012, Blundell et al., 2015, Polson et al., 2017, Rezende et al., 2014,**  
 40 **Li and Gal, 2017**].

41 This paper fills the gap in the literature by proposing a method called *Minimization by Incremental*  
 42 *Stochastic Surrogate Optimization (MISSO)*, designed for the nonconvex and nonsmooth finite sum  
 43 optimization, with a finite-time convergence guarantee. Our work aims at formulating a *generic*  
 44 *class* of incremental stochastic surrogate methods for nonconvex optimization and building the the-  
 45 ory to understand its behavior. In particular, we provide convergence guarantees for stochastic EM  
 46 and Variational Inference-type methods, under mild conditions. In summary, our contributions are:

- 47 • we propose a unifying framework of analysis for incremental stochastic surrogate optimiza-  
 48 tion when the surrogates are defined as expectations of tractable functions. The proposed  
 49 MISSO method is built on the Monte Carlo integration of the intractable surrogate function,  
 50 *i.e.*, a doubly stochastic surrogate optimization scheme.
- 51 • we present an incremental update of the commonly used variational inference and Monte  
 52 Carlo EM methods as special cases of our newly introduced framework. The analysis of  
 53 those two algorithms is thus conducted under this unifying framework of analysis.
- 54 • we establish both asymptotic and non-asymptotic convergence for the MISSO method. In  
 55 particular, the MISSO method converges almost surely to a stationary point and in  $\mathcal{O}(n/\epsilon)$   
 56 iterations to an  $\epsilon$ -stationary point, see Theorem 1.
- 57 • in essence, we relax the class of surrogate functions used in MISO [**Mairal, 2015**] and  
 58 allow for intractable surrogates that can only be evaluated by Monte-Carlo approximations.  
 59 Working at the crossroads of *Optimization* and *Sampling* constitutes what we believe to be  
 60 the novelty and the technicality of our framework and theoretical results.

61 In Section 2, we review the techniques for incremental minimization of finite sum functions based  
 62 on the MM principle; specifically, we review the MISO method [**Mairal, 2015**], and present a class  
 63 of surrogate functions expressed as an expectation over a latent space. The MISSO method is then  
 64 introduced for the latter class of intractable surrogate functions requiring approximation. In Sec-  
 65 tion 3, we provide the asymptotic and non-asymptotic convergence analysis for the MISSO method  
 66 (and of the MISO [**Mairal, 2015**] one as a special case). Section 4 presents numerical applications  
 67 including parameter inference for logistic regression with missing data and variational inference for  
 68 two types of Bayesian neural networks. The proofs of theoretical results are reported as Supplement.

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

$$f'(\theta, d) := \lim_{t \rightarrow 0^+} \frac{f(\theta + td) - f(\theta)}{t} . \quad (2)$$

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

## 73 2 Incremental Minimization of Finite Sum Nonconvex Functions

74 The objective function in (1) is composed of a finite sum of possibly nonsmooth and nonconvex  
 75 functions. A popular approach here is to apply the MM method, which tackles (1) through alter-  
 76 nating between two steps — (i) minimizing a *surrogate* function which upper bounds the original  
 77 objective function; and (ii) updating the surrogate function to tighten the upper bound.

78 As mentioned in the introduction, the MISO method [**Mairal, 2015**] is developed as an iterative  
 79 scheme that only updates the surrogate functions *partially* at each iteration. Formally, for any  $i \in$   
 80  $\llbracket 1, n \rrbracket$ , we consider a surrogate function  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  which satisfies the assumptions **(H1, H2)**:

81 **H1.** For all  $i \in \llbracket 1, n \rrbracket$  and  $\bar{\theta} \in \Theta$ ,  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  is convex w.r.t.  $\theta$ , and it holds

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) \geq \mathcal{L}_i(\theta), \quad \forall \theta \in \Theta , \quad (3)$$

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

83 **H2.** For any  $\bar{\theta}_i \in \Theta$ ,  $i \in \llbracket 1, n \rrbracket$  and some  $\epsilon > 0$ , the difference function  $\hat{e}(\theta; \{\bar{\theta}_i\}_{i=1}^n) :=$   
84  $\frac{1}{n} \sum_{i=1}^n \hat{\mathcal{L}}_i(\theta; \bar{\theta}_i) - \mathcal{L}(\theta)$  is defined for all  $\theta \in \Theta_\epsilon$  and differentiable for all  $\theta \in \Theta$ , where  
85  $\Theta_\epsilon = \{\theta \in \mathbb{R}^d, \inf_{\theta' \in \Theta} \|\theta - \theta'\| < \epsilon\}$  is an  $\epsilon$ -neighborhood set of  $\Theta$ . Moreover, for some constant  
86  $L$ , the gradient satisfies

$$\|\nabla \hat{e}(\theta; \{\bar{\theta}_i\}_{i=1}^n)\|^2 \leq 2L \hat{e}(\theta; \{\bar{\theta}_i\}_{i=1}^n), \forall \theta \in \Theta. \quad (4)$$

87 We remark that H1 is a common assumption  
88 used for surrogate functions, see [Mairal, 2015,  
89 Section 2.3]. H2 can be satisfied when the differ-  
90 ence function  $\hat{e}(\theta; \{\bar{\theta}_i\}_{i=1}^n)$  is  $L$ -smooth, i.e.,  $\hat{e}$   
91 is differentiable on  $\Theta$  and its gradient  $\nabla \hat{e}$  is  $L$ -  
92 Lipschitz,  $\forall \theta \in \Theta$ . H2 can be implied by apply-  
93 ing [Razaviyayn et al., 2013, Proposition 1].

94 The inequality (3) implies  $\hat{\mathcal{L}}_i(\theta; \bar{\theta}) \geq \mathcal{L}_i(\theta) >$   
95  $-\infty$  for any  $\theta \in \Theta$ . The MISO method is  
96 an incremental version of the MM method, as  
97 summarized by Algorithm 1, which shows that  
98 the MISO method maintains an iteratively up-  
99 dated set of upper-bounding surrogate functions  
100  $\{\mathcal{A}_i^k(\theta)\}_{i=1}^n$  and updates the iterate via minimiz-  
101 ing the average of the surrogate functions.

102 Particularly, only one out of the  $n$  surrogate functions is updated at each iteration [cf. Line 5] and  
103 the sum function  $\frac{1}{n} \sum_{i=1}^n \mathcal{A}_i^{k+1}(\theta)$  is designed to be ‘easy to optimize’, which, for example, can be  
104 a sum of quadratic functions. As such, the MISO method is suitable for large-scale optimization as  
105 the computation cost per iteration is independent of  $n$ . Under H1, H2, it was shown that the MISO  
106 method converges almost surely to a stationary point of (1) [Mairal, 2015, Prop. 3.1].

107 We now consider the case when the surrogate functions  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  are intractable. Let  $Z$  be a mea-  
108 surable set,  $p_i : Z \times \Theta \rightarrow \mathbb{R}_+$  a probability density function,  $r_i : \Theta \times \Theta \times Z \rightarrow \mathbb{R}$  a measurable  
109 function and  $\mu_i$  a  $\sigma$ -finite measure. We consider surrogate functions which satisfy H1, H2 and that  
110 can be expressed as an expectation, i.e.:

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) := \int_Z r_i(\theta; \bar{\theta}, z_i) p_i(z_i; \bar{\theta}) \mu_i(dz_i) \quad \forall (\theta, \bar{\theta}) \in \Theta \times \Theta. \quad (5)$$

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

113 In this paper, we propose the *Minimization by Incremental Stochastic Surrogate Optimization*  
114 (MISSO) method which replaces the expectation in (5) by *Monte Carlo* integration and then op-  
115 timizes the objective function (1) in an incremental manner. Denote by  $M \in \mathbb{N}$  the Monte Carlo  
116 batch size and let  $\{z_m \in Z\}_{m=1}^M$  be a set of samples. These samples can be drawn (Case 1) i.i.d.  
117 from the distribution  $p_i(\cdot; \bar{\theta})$  or (Case 2) from a Markov chain with stationary distribution  $p_i(\cdot; \bar{\theta})$ ;  
118 see Section 3 for illustrations. To this end, we define the stochastic surrogate as follows:

$$\tilde{\mathcal{L}}_i(\theta; \bar{\theta}, \{z_m\}_{m=1}^M) := \frac{1}{M} \sum_{m=1}^M r_i(\theta; \bar{\theta}, z_m), \quad (6)$$

119 and we summarize the proposed MISSO method in Algorithm 2. Compared to the MISO method,  
120 there is a crucial difference in that the MISSO method involves two types of randomness. The first  
121 level of randomness comes from the selection of  $i_k$  in Line 5. The second level of randomness stems  
122 from the set of Monte Carlo approximated functions  $\tilde{\mathcal{A}}_i^k(\theta)$  used in lieu of  $\mathcal{A}_i^k(\theta)$  in Line 6 when  
123 optimizing for the next iterate  $\theta^{(k)}$ . We now discuss two applications of the MISSO method.

124 **Example 1: Maximum Likelihood Estimation for Latent Variable Model.** Latent variable mod-  
125 els [Bishop, 2006] are constructed by introducing unobserved (latent) variables which help explain  
126 the observed data. We consider  $n$  independent observations  $((y_i, z_i), i \in \llbracket n \rrbracket)$  where  $y_i$  is observed

---

**Algorithm 1** The MISO method [Mairal, 2015].

---

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

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

- 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** The 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

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

- 4: **for**  $k = 0, 1, \dots, K_{\max}$  **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  $p_i(\cdot; \theta^{(k)})$ .
- 7:   Update the individual surrogate functions recursively as:

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

- 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**
- 

127 and  $z_i$  is latent. In this incomplete data framework, define  $\{f_i(z_i, \theta), \theta \in \Theta\}$  to be the complete  
 128 data likelihood models, *i.e.*, the joint likelihood of the observations and latent variables. Let

$$g_i(\theta) := \int_{\mathcal{Z}} f_i(z_i, \theta) \mu_i(dz_i), \quad i \in \llbracket 1, n \rrbracket, \quad \theta \in \Theta$$

129 denote the incomplete data likelihood, *i.e.*, the marginal likelihood of the observations  $y_i$ . For ease  
 130 of notations, the dependence on the observations is made implicit. The maximum likelihood (ML)  
 131 estimation problem sets the individual objective function  $\mathcal{L}_i(\theta)$  to be the  $i$ -th negated incomplete  
 132 data log-likelihood  $\mathcal{L}_i(\theta) := -\log g_i(\theta)$ .

133 Assume, without loss of generality, that  $g_i(\theta) \neq 0$  for all  $\theta \in \Theta$ . We define by  $p_i(z_i, \theta) :=$   
 134  $f_i(z_i, \theta)/g_i(\theta)$  the conditional distribution of the latent variable  $z_i$  given the observations  $y_i$ . A sur-  
 135 rogate function  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  satisfying H1 can be obtained through writing  $f_i(z_i, \theta) = \frac{f_i(z_i, \theta)}{p_i(z_i, \bar{\theta})} p_i(z_i, \bar{\theta})$   
 136 and applying the Jensen inequality:

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) = \int_{\mathcal{Z}} \underbrace{\log(p_i(z_i, \bar{\theta})/f_i(z_i, \theta))}_{=r_i(\theta; \bar{\theta}, z_i)} p_i(z_i, \bar{\theta}) \mu_i(dz_i). \quad (7)$$

137 We note that H2 can also be verified for common distribution models. We can apply the MISSO  
 138 method following the above specification of  $r_i(\theta; \bar{\theta}, z_i)$  and  $p_i(z_i, \bar{\theta})$ .

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

$$p(y, w|x) = \pi(w) \prod_{i=1}^n p(y_i|x_i, w). \quad (8)$$

142 Our goal is to compute the posterior distribution  $p(w|y, x)$ . In most cases, the posterior dis-  
 143 tribution  $p(w|y, x)$  is intractable and is approximated using a family of parametric distributions,  
 144  $\{q(w, \theta), \theta \in \Theta\}$ . The variational inference (VI) problem [Blei et al., 2017] boils down to minimiz-  
 145 ing the Kullback-Leibler (KL) divergence between  $q(w, \theta)$  and the posterior distribution  $p(w|y, x)$ :

$$\min_{\theta \in \Theta} \mathcal{L}(\theta) := \text{KL}(q(w; \theta) || p(w|y, x)) := \mathbb{E}_{q(w; \theta)} [\log(q(w; \theta)/p(w|y, x))] . \quad (9)$$

147 Using (8), we decompose  $\mathcal{L}(\theta) = n^{-1} \sum_{i=1}^n \mathcal{L}_i(\theta) + \text{const.}$  where:

$$\mathcal{L}_i(\theta) := -\mathbb{E}_{q(w; \theta)} [\log p(y_i|x_i, w)] + \frac{1}{n} \mathbb{E}_{q(w; \theta)} [\log q(w; \theta)/\pi(w)] := r_i(\theta) + d(\theta). \quad (10)$$

148 Directly optimizing the finite sum objective function in (9) can be difficult. First, with  $n \gg 1$ ,  
 149 evaluating the objective function  $\mathcal{L}(\theta)$  requires a full pass over the entire dataset. Second, for some

complex models, the expectations in (10) can be intractable even if we assume a simple parametric model for  $q(w; \theta)$ . Assume that  $\mathcal{L}_i$  is L-smooth. We apply the MISSO method with a quadratic surrogate function defined as:

$$\hat{\mathcal{L}}_i(\theta; \bar{\theta}) := \mathcal{L}_i(\bar{\theta}) + \langle \nabla_{\theta} \mathcal{L}_i(\bar{\theta}) | \theta - \bar{\theta} \rangle + \frac{L}{2} \|\theta - \bar{\theta}\|^2, \quad (\theta, \bar{\theta}) \in \Theta^2. \quad (11)$$

It is easily checked that the quadratic function  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  satisfies H1, H2. To compute the gradient  $\nabla \mathcal{L}_i(\bar{\theta})$ , 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.  $\theta \in \Theta$  which is designed such that the law of  $w = t(z, \bar{\theta})$  is  $q(\cdot, \bar{\theta})$ , where  $z \sim \mathcal{N}_d(0, \mathbf{I})$ . By [Blundell et al., 2015, Proposition 1], the gradient of  $-r_i(\cdot)$  in (10) is:

$$\nabla_{\theta} \mathbb{E}_{q(w; \bar{\theta})} [\log p(y_i | x_i, w)] = \mathbb{E}_{z \sim \mathcal{N}_d(0, \mathbf{I})} [J_{\theta}^t(z, \bar{\theta}) \nabla_w \log p(y_i | x_i, w) |_{w=t(z, \bar{\theta})}], \quad (12)$$

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 as the gradient of the KL between the prior distribution  $\pi(\cdot)$  and the variational candidate  $q(\cdot, \bar{\theta})$ .

$$r_i(\theta; \bar{\theta}, z) := \langle \nabla_{\theta} d(\bar{\theta}) - J_{\theta}^t(z, \bar{\theta}) \nabla_w \log p(y_i | x_i, w) |_{w=t(z, \bar{\theta})} | \theta - \bar{\theta} \rangle + \frac{L}{2} \|\theta - \bar{\theta}\|^2. \quad (13)$$

Finally, using (11) and (13), the surrogate function (6) 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  $\{z_m\}_{m=1}^M$  are i.i.d samples drawn from  $\mathcal{N}(0, \mathbf{I})$ .

### 3 Convergence Analysis

We now provide asymptotic and non-asymptotic convergence results of our method. Assume:

**H3.** For all  $i \in \llbracket 1, n \rrbracket$ ,  $\bar{\theta} \in \Theta$ ,  $z_i \in \mathcal{Z}$ ,  $r_i(\cdot; \bar{\theta}, z_i)$  is convex on  $\Theta$  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 MC approximation, introduced in (6), as:

**H4.** For the samples  $\{z_{i,m}\}_{m=1}^M$ , there exist finite constants  $C_r$  and  $C_{gr}$  such that

$$C_r := \sup_{\bar{\theta} \in \Theta} \sup_{M > 0} \frac{1}{\sqrt{M}} \mathbb{E}_{\bar{\theta}} \left[ \sup_{\theta \in \Theta} \left| \sum_{m=1}^M \{r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta})\} \right| \right]$$

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

for all  $i \in \llbracket 1, n \rrbracket$ , 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  $\Pi_{i, \bar{\theta}}$ , and stationary distribution  $p_i(\cdot; \bar{\theta})$ .

**Some intuitions behind the controlling terms:** It is 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 controlled random variable is an image of a continuous function defined as  $r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta})$  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 bracketing number) as developed in [Van der Vaart, 2000, Vershynin, 2018, Wainwright, 2019]. A collection of results from those references gives intuition behind our assumption H4, which is classical in empirical processes. In [Vershynin, 2018, Theorem 8.2.3], the authors recall the uniform law of large numbers:

$$\mathbb{E} \left[ \sup_{f \in \mathcal{F}} \left| \frac{1}{M} \sum_{i=1}^M f(z_{i,m}) - \mathbb{E}[f(z_i)] \right| \right] \leq \frac{CL}{\sqrt{M}} \quad \text{for all } z_{i,m}, i \in \llbracket 1, M \rrbracket,$$

where  $\mathcal{F}$  is a class of  $L$ -Lipschitz functions. Moreover, in [Vershynin, 2018, Theorem 8.1.3] and [Wainwright, 2019, Theorem 5.22], the application of the Dudley inequality yields:

$$\mathbb{E}[\sup_{f \in \mathcal{F}} |X_f - X_0|] \leq \frac{1}{\sqrt{M}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_{\infty}, \varepsilon)} d\varepsilon,$$

where  $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon)$  is the bracketing number and  $\epsilon$  denotes the level of approximation (the bracketing number goes to infinity when  $\epsilon \rightarrow 0$ ). Finally, in [Van der Vaart, 2000, p.271, Example],  $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon)$  is bounded from above for a class of parametric functions  $\mathcal{F} = f_\theta : \theta \in \Theta$ :

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \epsilon) \leq K \left( \frac{\text{diam } \Theta}{\epsilon} \right)^d, \quad \text{for all } 0 < \epsilon < \text{diam } \Theta.$$

The authors acknowledge that those bounds are a dramatic manifestation of the curse of dimensionality happening when sampling is needed. Nevertheless, the dependence on the dimension highly depends on the class of surrogate functions  $\mathcal{F}$  used in our scheme, as smaller bounds on these controlling terms can be derived for simpler class of functions, such as quadratic functions.

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

$$g(\bar{\theta}) := \inf_{\theta \in \Theta} \frac{\mathcal{L}'(\bar{\theta}, \theta - \bar{\theta})}{\|\bar{\theta} - \theta\|} \quad \text{and} \quad g(\bar{\theta}) = g_+(\bar{\theta}) - g_-(\bar{\theta}), \quad (14)$$

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]. Furthermore, suppose that the sequence  $\{\theta^{(k)}\}_{k \geq 0}$  has a limit point  $\bar{\theta}$  that is a stationary point, then one has  $\lim_{k \rightarrow \infty} g_-(\theta^{(k)}) = 0$ . Thus, the sequence  $\{\theta^{(k)}\}_{k \geq 0}$  is said to satisfy an *asymptotic stationary point condition*. This is equivalent to [Mairal, 2015, Definition 2.4].

To facilitate our analysis, we define  $\tau_i^k$  as the iteration index where the  $i$ -th function is last accessed in the MISSO method prior to iteration  $k$ ,  $\tau_{i_k}^{k+1} = k$  for instance. We define:

$$\widehat{\mathcal{L}}^{(k)}(\theta) := \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}_i(\theta; \theta^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\theta) := \widehat{\mathcal{L}}^{(k)}(\theta) - \mathcal{L}(\theta), \quad \overline{M}_{(k)} := \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}. \quad (15)$$

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

**Theorem 1.** Under H1-H4. 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:

$$\Delta_{(K_{\max})} := 2nL\mathbb{E}[\widehat{\mathcal{L}}^{(0)}(\theta^{(0)}) - \widehat{\mathcal{L}}^{(K_{\max})}(\theta^{(K_{\max})})] + 4LC_r \overline{M}_{(k)}.$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\theta^{(K)})\|^2] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \quad \text{and} \quad \mathbb{E}[g_-(\theta^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\text{gr}}}{K_{\max}} \overline{M}_{(k)}. \quad (16)$$

Note that  $\Delta_{(K_{\max})}$  is finite for any  $K_{\max} \in \mathbb{N}$ .

**Iteration Complexity of MISSO.** As expected, the MISSO method converges to a stationary point of (1) asymptotically and at a sublinear rate  $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{\Delta_{(K_{\max})}/K_{\max}})$ . In other terms, MISSO requires  $\mathcal{O}(nL/\epsilon)$  iterations to reach an  $\epsilon$ -stationary point when the suboptimality condition, that characterizes stationarity, is  $\mathbb{E}[\|g_-(\theta^{(K)})\|^2]$ . Note that this stationarity criterion are similar to the usual quantity used in stochastic nonconvex optimization, i.e.,  $\mathbb{E}[\|\nabla \mathcal{L}(\theta^{(K)})\|^2]$ . In fact, when the optimization problem (1) is unconstrained, i.e.,  $\Theta = \mathbb{R}^p$ , then  $\mathbb{E}[g(\theta^{(K)})] = \mathbb{E}[\nabla \mathcal{L}(\theta^{(K)})]$ .

**Sample Complexity of MISSO.** Regarding the sample complexity of our method, setting  $M_{(k)} = k^2/n^2$ , as a non-decreasing sequence of integers satisfying  $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$ , in order to keep  $\Delta_{(K_{\max})} \asymp nL$ , then the MISSO method requires  $\sum_{k=0}^{nL/\epsilon} k^2/n^2 = nL^3/\epsilon^3$  samples to reach an  $\epsilon$ -stationary point.

Furthermore, we remark that the MISSO 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. H4], Eq. (16) 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 H1-H4. 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 a.s. to zero, i.e.,  $\lim_{k \rightarrow \infty} g_{-}(\theta^{(k)}) \stackrel{a.s.}{=} 0$ .
2. the objective value  $\mathcal{L}(\theta^{(k)})$  converges a.s. to a finite number  $\underline{\mathcal{L}}$ , i.e.,  $\lim_{k \rightarrow \infty} \mathcal{L}(\theta^{(k)}) \stackrel{a.s.}{=} \underline{\mathcal{L}}$ .

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  $z_i = (z_{i,j}, j \in \llbracket p \rrbracket)$  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:

$$p_i(y_i | z_i) = S(\delta^\top \bar{z}_i)^{y_i} (1 - S(\delta^\top \bar{z}_i))^{1-y_i}, \quad (17)$$

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

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

**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. This dataset includes information from the first stage of the trauma, namely initial observations on the patient's accident site to the last stage being intense care at the hospital and counts more than 200 variables measured for more than 7 000 patients. Since the dataset considered is heterogeneous – coming from multiple sources with frequently missed entries – we apply the latent data model described in (17) to *predict the risk of a severe hemorrhage* which is one of the main cause of death after a major trauma.

Similar to [Jiang et al., 2018], we select  $p = 16$  influential quantitative measurements, on  $n = 6384$  patients. For the Monte Carlo sampling of  $z_{i,\text{mis}}$ , required while running MISSO, we run a Metropolis-Hastings algorithm with the target distribution  $p(\cdot | z_{i,\text{obs}}, y_i; \theta^{(k)})$ .

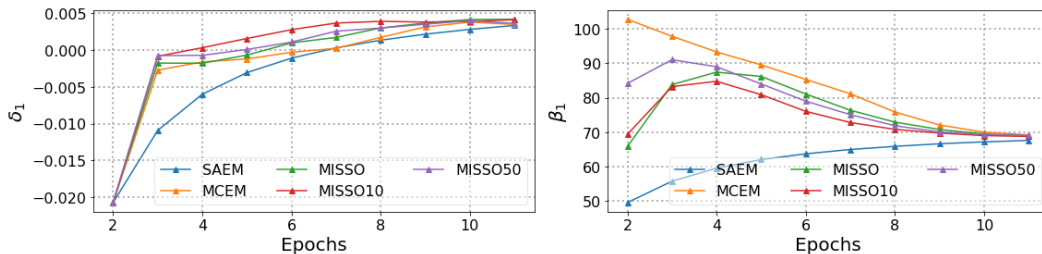


Figure 1: Convergence of parameters  $\delta$  and  $\beta$  for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against number of passes over the data.

We compare in Figure 1 the convergence behavior of the estimated parameters  $\delta$  and  $\beta$  using SAEM [Delyon et al., 1999] (with stepsize  $\gamma_k = 1/k^\alpha$  where  $\alpha = 0.6$  after tuning), 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

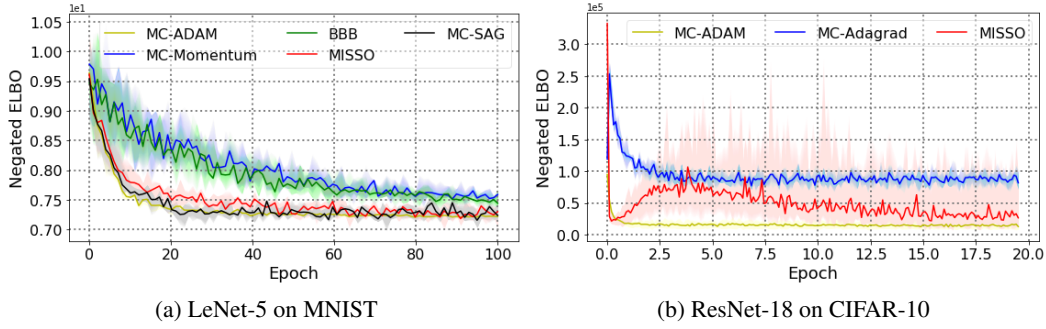


Figure 2: Negated ELBO versus epochs elapsed for fitting (a) Bayesian LeNet-5 on MNIST and (b) Bayesian ResNet-18 on CIFAR-10. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

method – the default settings with 1 (MISSO), 10% (MISSO10) and 50% (MISSO50) minibatches 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. Though wall clock times are similar for all methods, they are reported in the appendix for completeness.

#### 4.2 Training Bayesian CNN using MISSO

This application follows **Example 2** described in Section 2. We use variational inference and the ELBO loss (10) to fit Bayesian Neural Networks on different datasets. At iteration  $k$ , minimizing the sum of stochastic surrogates defined as in (6) and (13) 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, with  $\tilde{w} = t(\theta^{(k-1)}, z_m^{(k)})$ , as

$$\begin{aligned}\mu_\ell^{(k)} &= \hat{\mu}_\ell^{(\tau^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\mu_\ell, i}^{(k)}, \\ \hat{\delta}_{\mu_\ell, i_k}^{(k)} &= -\frac{1}{M(k)} \sum_{m=1}^{M(k)} \nabla_w \log p(y_{i_k} | x_{i_k}, \tilde{w}) + \nabla_{\mu_\ell} d(\theta^{(k-1)}),\end{aligned}$$

where  $\hat{\mu}_\ell^{(\tau^k)} = \frac{1}{n} \sum_{i=1}^n \mu_\ell^{(\tau_i^k)}$  and  $d(\theta) = n^{-1} \sum_{\ell=1}^d (-\log(\sigma) + (\sigma^2 + \mu_\ell^2)/2 - 1/2)$ .

**Bayesian LeNet-5 on MNIST [LeCun et al., 1998]:** We apply the MISSO method to fit a Bayesian variant of LeNet-5 [LeCun et al., 1998]. 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 (8), the weights are assumed independent and identically distributed according to  $\mathcal{N}(0, 1)$ . We also assume that  $q(\cdot; \theta) \equiv \mathcal{N}(\mu, \sigma^2 \mathbf{I})$ . The variational posterior parameters are thus  $\theta = (\mu, \sigma)$  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  $w = t(\theta, z) = \mu + \sigma z$  with  $z \sim \mathcal{N}(0, \mathbf{I})$ .

**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 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 6 000 images per class. As in the previous example, the weights are assumed independent and identically distributed according to  $\mathcal{N}(0, \mathbf{I})$ . Standard hyperparameters values found in the literature, such as the annealing constant or the number of MC samples, were used for the benchmark methods. For efficiency purpose and lower variance, the Flipout estimator [Wen et al., 2018] is used.

**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 Back-prop* (BBB) [Blundell et al., 2015] and our proposed MISSO method. For all these methods, the



loss function (10) and its gradients were computed by Monte Carlo integration based on the re-parametrization described above. The mini-batch of indices and MC samples are respectively set to 128 and  $M_{(k)} = k$ . The learning rates are set to  $10^{-3}$  for LeNet-5 and  $10^{-4}$  for Resnet-18.

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 lower training loss (and also faster). Our MISSO method is similar to the Monte Carlo variant of ADAM but slower than Adagrad optimizer. 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 baselines methods. We report wall clock times for all methods in the appendix for completeness.

## 5 Conclusion

We present a unifying framework for minimizing a nonconvex and nonsmooth 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 nonconvex 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 Bayesian variants of two Convolutional Neural Networks on benchmark datasets.

## References

- Christopher M Bishop. *Pattern recognition and machine learning*. springer, 2006.
- David M. Blei, Alp Kucukelbir, and Jon 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>.
- Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In *International Conference on Machine Learning*, pages 1613–1622, 2015.
- Bernard Delyon, Marc Lavielle, and Eric 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>.
- Roger Fletcher, Nicholas IM Gould, Sven Leyffer, Philippe L Toint, and Andreas 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.
- Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.
- Wei Jiang, Julie Josse, and Marc Lavielle. Logistic regression with missing covariates–parameter estimation, model selection and prediction. 2018.
- Michael I. Jordan, Zoubin Ghahramani, Tommi S. Jaakkola, and Lawrence K. Saul. An introduction to variational methods for graphical models. *Mach. Learn.*, 37(2):183–233, November 1999. ISSN 0885-6125. doi: 10.1023/A:1007665907178. URL <https://doi.org/10.1023/A:1007665907178>.
- Diederik P. Kingma and Jimmy 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>.
- Diederik P. Kingma and Max 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>.
- Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- Kenneth Lange. *MM Optimization Algorithms*. SIAM-Society for Industrial and Applied Mathematics, USA, 2016. ISBN 1611974399, 9781611974393.
- Yann LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- Yann LeCun, Léon Bottou, Yoshua Bengio, Patrick Haffner, et al. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- Yingzhen Li and Yarin 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.
- Julien 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>.

- Geoffrey J. McLachlan and Thriyambakam 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>.
- Sean P Meyn and Richard L Tweedie. *Markov chains and stochastic stability*. Springer Science & Business Media, 2012.
- Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.
- J.W. Paisley, D.M. Blei, and M.I. Jordan. Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress, 2012.
- Nicholas G Polson, Vadim Sokolov, et al. Deep learning: a bayesian perspective. *Bayesian Analysis*, 12(4):1275–1304, 2017.
- Xun Qian, Alibek Sailanbayev, Konstantin Mishchenko, and Peter Richtárik. Miso is making a comeback with better proofs and rates. *arXiv preprint arXiv:1906.01474*, 2019.
- Meisam Razaviyayn, Mingyi Hong, and Zhi-Quan Luo. A unified convergence analysis of block successive minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–1153, 2013.
- Danilo Jimenez Rezende, Shakir Mohamed, and Daan Wierstra. Stochastic backpropagation and approximate inference in deep generative models. In *International Conference on Machine Learning*, pages 1278–1286, 2014.
- Mark Schmidt, Nicolas Le Roux, and Francis Bach. Minimizing finite sums with the stochastic average gradient. *Mathematical Programming*, 162(1-2):83–112, 2017.
- Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.
- Martin J Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge University Press, 2019.
- Greg C. G. Wei and Martin 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>.
- Yeming Wen, Paul Vicol, Jimmy Ba, Dustin Tran, and Roger Grosse. Flipout: Efficient pseudo-independent weight perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.

## A Proofs of the Theoretical Results

### A.1 Proof of Theorem 1

**Theorem.** Under H1-H4. 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:

$$\Delta_{(K_{\max})} := 2nL\mathbb{E}[\tilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \tilde{\mathcal{L}}^{(K_{\max})}(\boldsymbol{\theta}^{(K_{\max})})] + 4LC_r\overline{M}_{(k)}.$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}[\|\nabla \tilde{\mathcal{L}}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \quad \text{and} \quad \mathbb{E}[g_{-}(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\text{gr}}}{K_{\max}}\overline{M}_{(k)}.$$

**Proof** We begin by recalling the definition

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

Notice that

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

Furthermore, we recall that

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

Due to H2, we have

$$\|\nabla \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\hat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \quad (18)$$

To prove the first bound in (16), using the optimality of  $\boldsymbol{\theta}^{(k+1)}$ , one has

$$\begin{aligned} \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) &\leq \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\ &= \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{n} (\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})). \end{aligned} \quad (19)$$

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

We observe that the conditional expectation evaluates to

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

where the last inequality is due to H4. Moreover,

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

Taking the conditional expectations on both sides of (19) and re-arranging terms give:

$$\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) | \mathcal{F}_k] + \frac{C_r}{\sqrt{M_{(k)}}}. \quad (20)$$





404 This concludes our proof for the first inequality in (16).

405 To prove the second inequality of (16), we define the shorthand notations  $g^{(k)} := g(\theta^{(k)})$ ,  $g_-^{(k)} :=$   
 406  $-\min\{0, g^{(k)}\}$ ,  $g_+^{(k)} := \max\{0, g^{(k)}\}$ . We observe that

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

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

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

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

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

412 We have

$$g^{(k)} \geq -\|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| + \inf_{\theta \in \Theta} (-\epsilon^{(k)}(\theta)) \geq -\|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| - \sup_{\theta \in \Theta} |\epsilon^{(k)}(\theta)|. \quad (25)$$

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

$$g_-^{(k)} \leq \|\nabla \widehat{e}^{(k)}(\theta^{(k)})\| + \sup_{\theta \in \Theta} |\epsilon^{(k)}(\theta)|. \quad (26)$$

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

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

416 We note that

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

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

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

418 where (a) is due to H4 and (b) is due to (24). This implies

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

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

## 420 A.2 Proof of Theorem 2

421 **Theorem.** Under H1-H4. In addition, assume that  $\{M_{(k)}\}_{k \geq 0}$  is a non-decreasing sequence of  
 422 integers which satisfies  $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$ . Then:

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

425 **Proof** We apply the following auxiliary lemma which proof can be found in Appendix A.3 for the  
 426 readability of the current proof:

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

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1} \quad (27)$$

430 then:

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

434 We proceed from (19) by re-arranging terms and observing that

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

435 Our idea is to apply Lemma 1. Under H1, the finite sum of surrogate functions  $\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta})$ , defined in  
 436 (15), is lower bounded by a constant  $c_k > -\infty$  for any  $\boldsymbol{\theta}$ . To this end, we observe that

$$V_k := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \inf_{k \geq 0} c_k \geq 0 \quad (28)$$

437 is a non-negative random variable.

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

$$X_k := \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(\tau_{i_k}^k)}; \boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \geq 0 . \quad (29)$$

439 Thirdly, we define

$$\begin{aligned} E_k &= -(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k, m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})) . \end{aligned} \quad (30)$$

440 Note that from the definitions (28), (29), (30), we have  $V_{k+1} \leq V_k - X_k + E_k$  for any  $k \geq 1$ .

441 Under H4, we observe that

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

442

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

443

$$\mathbb{E}[|\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})|] \leq \frac{1}{n} \sum_{i=1}^n C_r \mathbb{E}[M_{(\tau_i^k)}^{-1/2}]$$

444 Therefore,

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

445 Using (24) and the assumption on the sequence  $\{M_{(k)}\}_{k \geq 0}$ , we obtain that

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

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

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

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

$$\lim_{k \rightarrow \infty} \hat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0 \quad \text{a.s.} \quad (31)$$

449 and subsequently applying (18), we have  $\lim_{k \rightarrow \infty} \|\hat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| = 0$  almost surely. Finally, it follows  
 450 from (18) and (26) that

$$\lim_{k \rightarrow \infty} g_-^{(k)} \leq \lim_{k \rightarrow \infty} \sqrt{2L} \sqrt{\hat{e}^{(k)}(\boldsymbol{\theta}^{(k)})} + \lim_{k \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})| = 0, \quad (32)$$

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

453 Finally, we prove that  $\mathcal{L}(\boldsymbol{\theta}^{(k)})$  converges almost surely. As a consequence of Lemma 1, it is clear that  
 454  $\{V_k\}_{k \geq 0}$  converges almost surely and so is  $\{\hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\}_{k \geq 0}$ , i.e., we have  $\lim_{k \rightarrow \infty} \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$ .  
 455 Applying (31) implies that

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

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

### 457 A.3 Proof of Lemma 1

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

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

461 then:

462 (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  $V_{\infty}$ .

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

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

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

$$0 \leq V_k \leq V_0 - \sum_{j=1}^k X_j + \sum_{j=1}^k E_j \leq V_0 + \sum_{j=1}^k E_j, \quad (33)$$

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

467 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  
468  $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$ , the series  $\sum_{j=1}^{\infty} E_j$  converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \quad (34)$$

469 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:

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k]. \end{aligned} \quad (35)$$

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

$$\mathbb{E}[\liminf_{k \rightarrow \infty} |W_k|] = \mathbb{E}[|W_{\infty}|] \leq \liminf_{k \rightarrow \infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty, \quad (36)$$

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

475 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  
476  $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$  and by the monotone convergence theorem, we get:

$$\lim_{k \rightarrow \infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k \rightarrow \infty} U_k]. \quad (37)$$

477 Finally, we have:

$$\lim_{k \rightarrow \infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_{\infty} \quad \text{and} \quad \mathbb{E}[\lim_{k \rightarrow \infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_{\infty}]. \quad (38)$$

478 showing that  $\mathbb{E}[W_{\infty}] = w_{\infty}$  and concluding the proof of (ii). Moreover, using (35) we have that  
479  $W_k \leq W_{k-1} - X_k$  which yields:

$$\begin{aligned} \sum_{j=1}^{\infty} X_j &\leq W_0 - W_{\infty} < \infty, \\ \sum_{j=1}^{\infty} \mathbb{E}[X_j] &\leq \mathbb{E}[W_0] - w_{\infty} < \infty, \end{aligned} \quad (39)$$

480 an concludes the proof of the lemma.  $\square$

## 481 **B Practical Details for the Binary Logistic Regression on the Traumabase**

### 482 **B.1 Traumabase dataset quantitative variables**

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

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

## 491 B.2 Metropolis-Hastings algorithm

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

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

497 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  
 498 (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:  $z_{i,\text{mis},M}$ 

```

---

## 499 B.3 MISSO Update

500 **Choice of surrogate function for MISO:** We recall the MISO deterministic surrogate defined in  
 501 (7):

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

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

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

$$\begin{aligned}\hat{\mathcal{L}}_i(\theta; \bar{\theta}) &= - \int_{\mathcal{Z}} \log f_i(z_{i,\text{mis}}, z_{i,\text{obs}}, \theta) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) \\ &= - \int_{\mathcal{Z}} \log [p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) p_i(z_{i,\text{mis}}, \beta, \Omega)] p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}}) \\ &= \underbrace{- \int_{\mathcal{Z}} \log p_i(y_i | z_{i,\text{mis}}, z_{i,\text{obs}}, \delta) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}})}_{=\hat{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta})} - \underbrace{\int_{\mathcal{Z}} \log p_i(z_{i,\text{mis}}, \beta, \Omega) p_i(z_{i,\text{mis}}, \bar{\theta}) \mu_i(dz_{i,\text{mis}})}_{=\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})} .\end{aligned}\tag{40}$$



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

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

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

511 Recall that:

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

512 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  $p \times p$   
 513 matrix  $z_i z_i^\top$  is semi-definite positive we can assume that:

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

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

$$\bar{\mathcal{L}}_i^{(1)}(\delta, \bar{\theta}) = \hat{\mathcal{L}}_i^{(1)}(\bar{\delta}, \bar{\theta}) - D_i^\top (\delta - \bar{\delta}) + \frac{1}{2} (\delta - \bar{\delta}) H_i (\delta - \bar{\delta})^\top , \quad (41)$$

517 where:

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

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

$$\begin{aligned} \tilde{\mathcal{L}}^{(k)}(\theta) &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i(\theta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) \\ &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) - \frac{1}{n} \sum_{i=1}^n \tilde{D}_i^{(\tau_i^k)} (\delta - \delta^{(\tau_i^k)}) \\ &\quad + \frac{1}{2n} \sum_{i=1}^n (\delta - \delta^{(\tau_i^k)}) \left\{ \tilde{H}_i^{(\tau_i^k)} \right\} (\delta - \delta^{(\tau_i^k)})^\top , \end{aligned} \quad (42)$$

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

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

520 Minimizing the total surrogate (42) boils down to performing a quasi-Newton step. It is perhaps sen-  
 521 sible to apply some diagonal loading which is perfectly compatible with the surrogate interpretation  
 522 we just gave.

523 The logistic parameters are estimated as follows:

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

524 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 (41)  
 525 and which leads to the following quasi-Newton step:

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

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

527 **MISSO updates:** At the  $k$ -th iteration, and after the initialization, for all  $i \in \llbracket n \rrbracket$ , of the latent  
 528 variables ( $z_i^{(0)}$ ), the MISSO algorithm consists in picking an index  $i_k$  uniformly on  $\llbracket n \rrbracket$ , complet-  
 529 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  
 530 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  
 531 estimated parameters as follows:

$$\begin{aligned} \beta^{(k)} &= \arg \min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega^{(k)}, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{M(\tau_i^k)} \sum_{m=1}^{M(\tau_i^k)} z_{i,m}^{(k)}, \\ \Omega^{(k)} &= \arg \min_{\Omega \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta^{(k)}, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{M(\tau_i^k)} \sum_{m=1}^{M(\tau_i^k)} w_{i,m}^{(k)}, \\ \delta^{(k)} &= \frac{1}{n} \sum_{i=1}^n \delta^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1} \tilde{D}^{(k)}. \end{aligned} \quad (43)$$

532 where  $z_{i,m}^{(k)} = (z_{i, \text{mis}, m}^{(k)}, z_{i, \text{obs}})$  is composed of a simulated and an observed part,  $\tilde{D}^{(k)} =$   
 533  $\frac{1}{n} \sum_{i=1}^n \tilde{D}_i^{(\tau_i^k)}$ ,  $\tilde{H}^{(k)} = \frac{1}{n} \sum_{i=1}^n \tilde{H}_i^{(\tau_i^k)}$  and  $w_{i,m}^{(k)} = z_{i,m}^{(k)} (z_{i,m}^{(k)})^\top - \beta^{(k)} (\beta^{(k)})^\top$ . Be-  
 534 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  
 535  $\hat{\mathcal{L}}_i^{(1)}(\beta, \Omega, \bar{\theta})$  and  $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})$ , for all  $i \in \llbracket n \rrbracket$  as components of the surrogate function (40).

## 536 C Practical Details for the Incremental Variational Inference

### 537 C.1 Neural Networks Architecture

538 **Bayesian LeNet-5 Architecture:** We describe in Table 1 the architecture of the Convolutional  
 539 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

540 **Bayesian ResNet-18 Architecture:** We describe in Table 2 the architecture of the Resnet-18 we  
 541 train on CIFAR-10:

layer type	Output Size	ResNet-18	nonlinearity
conv1	$112 \times 112 \times 64$	$7 \times 7, 64$ , 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

## 542 C.2 Algorithms updates

543 First, we initialize the means  $\mu_\ell^{(0)}$  for  $\ell \in \llbracket d \rrbracket$  and variance estimates  $\sigma^{(0)}$ . At iteration  $k$ , minimizing  
544 the sum of stochastic surrogates defined as in (6) and (13) yields the following MISSO update —  
545 **step (i)** pick a function index  $i_k$  uniformly on  $\llbracket n \rrbracket$ ; **step (ii)** sample a Monte Carlo batch  $\{z_m^{(k)}\}_{m=1}^{M(k)}$   
546 from  $\mathcal{N}(0, \mathbf{I})$ ; and **step (iii)** update the parameters as

$$\mu_\ell^{(k)} = \frac{1}{n} \sum_{i=1}^n \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\mu_\ell, i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n} \sum_{i=1}^n \sigma^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\sigma, i}^{(k)}, \quad (44)$$

547 where we define the following gradient terms for all  $i \in \llbracket 1, n \rrbracket$ :

$$\begin{aligned} \hat{\delta}_{\mu_\ell, i}^{(k)} &= -\frac{1}{M(k)} \sum_{m=1}^{M(k)} \nabla_w \log p(y_i | x_i, w) \Big|_{w=t(\theta^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} d(\theta^{(k-1)}), \\ \hat{\delta}_{\sigma, i}^{(k)} &= -\frac{1}{M(k)} \sum_{m=1}^{M(k)} z_m^{(k)} \nabla_w \log p(y_i | x_i, w) \Big|_{w=t(\theta^{(k-1)}, z_m^{(k)})} + \nabla_{\sigma} d(\theta^{(k-1)}). \end{aligned} \quad (45)$$

548 Note that our analysis in the main text does require the parameter to be in a compact set. For the  
549 current estimation problem considered, this can be enforced in practice by restricting the parameters  
550 in a ball. In our simulation for the BNNs example, we did not implement the algorithms that stick  
551 closely to the compactness requirement for illustrative purposes. However, we observe empirically  
552 that the parameters are always bounded. The update rules can be easily modified to respect the  
553 requirement. For the considered VI problem, we recall the surrogate functions (11) are quadratic  
554 and indeed a simple projection step suffices to ensure boundedness of the iterates.

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

558 **Monte Carlo SAG update:** Set

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

559 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 (45) for  $i = i_k$ . The learning  
560 rate is set to  $\gamma = 10^{-3}$ .

561 **Bayes By Backprop update:** Set

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

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

563 **Monte Carlo Momentum update:** Set

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

564 where

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

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

566 **Monte Carlo ADAM update:** Set

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

567 where

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

568 and

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

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