

---

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

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 Many constrained, nonconvex and nonsmooth optimization problems can be tack-  
2 led using the majorization-minimization (MM) method which alternates between  
3 constructing a surrogate function which upper bounds the objective function, and  
4 then minimizing this surrogate. For problems which minimize a finite sum of  
5 functions, a stochastic version of the MM method selects a batch of functions  
6 at random at each iteration and optimizes the accumulated surrogate. However,  
7 in many cases of interest such as variational inference for latent variable mod-  
8 els, the surrogate functions are expressed as an expectation. In this contribution,  
9 we propose a doubly stochastic MM method based on Monte Carlo approxima-  
10 tion of these stochastic surrogates. We establish asymptotic and non-asymptotic  
11 convergence of our scheme in a constrained, nonconvex, nonsmooth optimization  
12 setting. We apply our new framework for inference of logistic regression model  
13 with missing data and for variational inference of Bayesian variants of LeNet-5  
14 and Resnet-18 on respectively the MNIST and CIFAR-10 datasets.

## 1 Introduction

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

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

17 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.

19 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].

31 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. For instance, this

is the case for inference in latent variable models via maximum likelihood [McLachlan and Krishnan, 2008]. 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 method called *Minimization by Incremental Stochastic Surrogate Optimization (MISSO)*, designed for the nonconvex and nonsmooth finite sum optimization, with a finite-time convergence guarantee. Our work aims at formulating a *generic class* of incremental stochastic surrogate methods for nonconvex optimization and building the theory to understand its behavior. In particular, we provide convergence guarantees for stochastic EM and Variational Inference-type methods, under mild conditions. In summary, our contributions are:

- we propose a unifying framework of analysis for incremental stochastic surrogate optimization when the surrogates are defined as 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 conducted 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, see Theorem 1.

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 [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). Section 4 presents numerical applications including parameter inference for logistic regression with missing data and variational inference for two types of Bayesian neural networks. The proofs of theoretical results are reported as Supplement.

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

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

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

## 2 Incremental Minimization of Finite Sum Nonconvex Functions

The objective function in (1) is composed of a finite sum of possibly nonsmooth and nonconvex functions. A popular approach here is to apply the MM method, which 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 [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  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  which satisfies the assumptions (H1, H2):

**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

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

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

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

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

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

The inequality (3) implies  $\hat{\mathcal{L}}_i(\theta; \bar{\theta}) \geq \mathcal{L}_i(\theta) > -\infty$  for any  $\theta \in \Theta$ . The MISO method is an incremental version of the MM method, as summarized by Algorithm 1, which shows that the MISO method maintains an iteratively updated set of upper-bounding surrogate functions  $\{\mathcal{A}_i^k(\theta)\}_{i=1}^n$  and updates the iterate via 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’, which, for example, 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  $n$ . Under H1, H2, it was shown that the MISO method converges almost surely to a stationary point of (1) [Mairal, 2015, Prop. 3.1].

We now consider the case when the surrogate functions  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  are intractable. Let  $Z$  be a measurable 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 function and  $\mu_i$  a  $\sigma$ -finite measure. We consider surrogate functions which satisfy H1, H2 and that can be expressed as an expectation, i.e.:

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

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

In this paper, we propose the *Minimization by Incremental Stochastic Surrogate Optimization* (MISSO) method which replaces the expectation in (5) by *Monte Carlo* integration and then optimizes the objective function (1) in an incremental manner. Denote by  $M \in \mathbb{N}$  the Monte Carlo 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. from the distribution  $p_i(\cdot; \bar{\theta})$  or (Case 2) from a Markov chain with stationary distribution  $p_i(\cdot; \bar{\theta})$ ; see Section 3 for illustrations. To this end, we define the stochastic surrogate as follows:

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

and we summarize the proposed MISSO method in Algorithm 2. Compared to the MISO method, there is a crucial difference in that the MISSO method involves two types of randomness. The first level of randomness comes from the selection of  $i_k$  in Line 5. The second level of randomness stems 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 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., the joint likelihood of the observations and latent variables. Let

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

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

---

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

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

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

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

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

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

135 Our goal is to compute the posterior distribution  $p(w|y, x)$ . In most cases, the posterior dis-  
 136 tribution  $p(w|y, x)$  is intractable and is approximated using a family of parametric distributions,  
 137  $\{q(w, \theta), \theta \in \Theta\}$ . The variational inference (VI) problem [Blei et al., 2017] boils down to minimiz-  
 138 ing the Kullback-Leibler (KL) divergence between  $q(w, \theta)$  and the posterior distribution  $p(w|y, x)$ :  
 139

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

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

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

141 Directly optimizing the finite sum objective function in (9) can be difficult. First, with  $n \gg 1$ ,  
 142 evaluating the objective function  $\mathcal{L}(\theta)$  requires a full pass over the entire dataset. Second, for some  
 143 complex models, the expectations in (10) can be intractable even if we assume a simple parametric  
 144 model for  $q(w; \theta)$ . Assume that  $\mathcal{L}_i$  is L-smooth. We apply the MISSO method with a quadratic  
 145 surrogate function defined as:

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

146 It is easily checked that the quadratic function  $\hat{\mathcal{L}}_i(\theta; \bar{\theta})$  satisfies H1, H2. To compute the gradient  
 147  $\nabla \mathcal{L}_i(\bar{\theta})$ , we apply the re-parametrization technique suggested in [Paisley et al., 2012, Kingma and  
 148 Welling, 2014, Blundell et al., 2015]. Let  $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$  be a differentiable function w.r.t.  $\theta \in \Theta$

149 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  
150 et al., 2015, Proposition 1], the gradient of  $-r_i(\cdot)$  in (10) is:

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

151 where for each  $z \in \mathbb{R}^d$ ,  $\mathbf{J}_{\theta}^t(z, \bar{\theta})$  is the Jacobian of the function  $t(z, \cdot)$  with respect to  $\theta$  evaluated at  
152  $\bar{\theta}$ . In addition, for most cases, the term  $\nabla d(\bar{\theta})$  can be evaluated in closed form as the gradient of the  
153 KL between the prior distribution  $\pi(\cdot)$  and the variational candidate  $q(\cdot, \theta)$ .

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

154 Finally, using (11) and (13), the surrogate function (6) is given by  $\tilde{\mathcal{L}}_i(\theta; \bar{\theta}, \{z_m\}_{m=1}^M) :=$   
155  $M^{-1} \sum_{m=1}^M r_i(\theta; \bar{\theta}, z_m)$  where  $\{z_m\}_{m=1}^M$  are i.i.d samples drawn from  $\mathcal{N}(0, \mathbf{I})$ .

### 156 3 Convergence Analysis

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

158 **H3.** For all  $i \in \llbracket 1, n \rrbracket$ ,  $\bar{\theta} \in \Theta$ ,  $z_i \in \mathbf{Z}$ ,  $r_i(\cdot; \bar{\theta}, z_i)$  is convex on  $\Theta$  and is lower bounded.

159 We are particularly interested in the *constrained optimization* setting where  $\Theta$  is a bounded set. To  
160 this end, we control the supremum norm of the MC approximation, introduced in (6), as:

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

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

162

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

163 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  
164 initial distribution  $\xi_i(\cdot; \bar{\theta})$ , transition kernel  $\Pi_{i, \bar{\theta}}$ , and stationary distribution  $p_i(\cdot; \bar{\theta})$ .

165 **Some intuitions behind the controlling terms:** It is common in statistical and optimization prob-  
166 lems, to deal with the manipulation and the control of random variables indexed by sets with an  
167 infinite number of elements. Here, the controlled random variable is an image of a continuous func-  
168 tion defined as  $r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta})$  for all  $z \in \mathbf{Z}$  and for fixed  $(\theta, \bar{\theta}) \in \Theta^2$ . To characterize  
169 such control, we will have recourse to the notion of metric entropy (or bracketing number) as de-  
170 veloped in [Van der Vaart, 2000, Vershynin, 2018, Wainwright, 2019]. A collection of results from  
171 those references gives intuition behind our assumption H4, which is classical in empirical processes.  
172 In [Vershynin, 2018, Theorem 8.2.3], the authors recall the uniform law of large numbers:

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

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

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

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

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

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

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

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

184 where  $g_+(\bar{\theta}) := \max\{0, g(\bar{\theta})\}$ ,  $g_-(\bar{\theta}) := -\min\{0, g(\bar{\theta})\}$  denote the positive and negative part of  
 185  $g(\bar{\theta})$ , respectively. Note that  $\bar{\theta}$  is a stationary point if and only if  $g_-(\bar{\theta}) = 0$  [Fletcher et al., 2002].  
 186 Furthermore, suppose that the sequence  $\{\theta^{(k)}\}_{k \geq 0}$  has a limit point  $\bar{\theta}$  that is a stationary point,  
 187 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*  
 188 *stationary point condition*. This is equivalent to [Mairal, 2015, Definition 2.4].

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

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

191 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:*

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

*Then we have following non-asymptotic bounds:*

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

192 Note that  $\Delta_{(K_{\max})}$  is finite for any  $K_{\max} \in \mathbb{N}$ . As expected, the MISSO method converges to a sta-  
 193 tionary point of (1) asymptotically and at a sublinear rate  $\mathbb{E}[g_-(\theta^{(K)})] \leq \mathcal{O}(\sqrt{1/K_{\max}})$ . Furthermore, we  
 194 remark that the MISO method can be analyzed in Theorem 1 as a special case of the MISSO method  
 195 satisfying  $C_r = C_{\text{gr}} = 0$ . In this case, while the asymptotic convergence is well known from [Mairal,  
 196 2015] [cf. H4], Eq. (16) gives a non-asymptotic rate of  $\mathbb{E}[g_-(\theta^{(K)})] \leq \mathcal{O}(\sqrt{nL/K_{\max}})$  which is new to  
 197 our best knowledge. Next, we show that under an additional assumption on the sequence of batch  
 198 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.,  $\lim_{k \rightarrow \infty} \mathcal{L}(\theta^{(k)}) \stackrel{a.s.}{=} \underline{\mathcal{L}}$ .*

199 In particular, the first result above shows that the sequence  $\{\theta^{(k)}\}_{k \geq 0}$  produced by the MISSO  
 200 method satisfies an *asymptotic stationary point condition*.

## 201 4 Numerical Experiments

### 202 4.1 Binary logistic regression with missing values

203 This application follows **Example 1** described in Section 2. We consider a binary regression setup,  
 204  $((y_i, z_i), i \in \llbracket n \rrbracket)$  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:

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

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  $p(\cdot|z_{i,\text{obs}}, y_i; \theta^{(k)})$ .

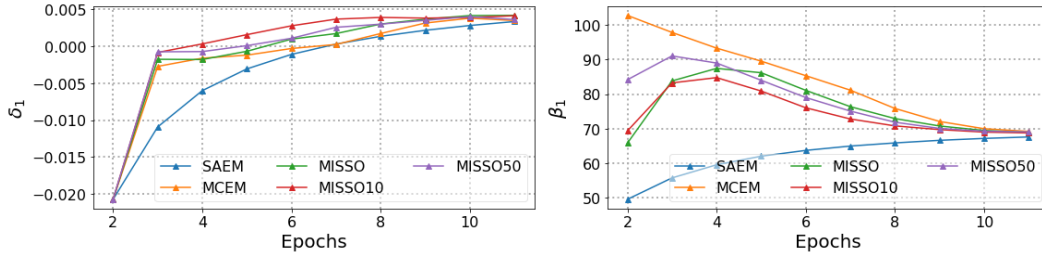


Figure 1: Convergence of first component of the vector of parameters  $\delta$  and  $\beta$  for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against No. of passes over the data.

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

## 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

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

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

238 **Bayesian LeNet-5 on MNIST [LeCun et al., 1998]:** We apply the MISSO method to fit a Bayesian  
 239 variant of LeNet-5 [LeCun et al., 1998]. We train this network on the MNIST dataset [LeCun,  
 240 1998]. The training set is composed of  $n = 55\,000$  handwritten digits,  $28 \times 28$  images. Each  
 241 image is labelled with its corresponding number (from zero to nine). Under the prior distribution  
 242  $\pi$ , see (8), the weights are assumed independent and identically distributed according to  $\mathcal{N}(0, 1)$ .  
 243 We also assume that  $q(\cdot; \theta) \equiv \mathcal{N}(\mu, \sigma^2 \mathbf{I})$ . The variational posterior parameters are thus  $\theta = (\mu, \sigma)$   
 244 where  $\mu = (\mu_\ell, \ell \in \llbracket d \rrbracket)$  where  $d$  is the number of weights in the neural network. We use the  
 245 re-parametrization as  $w = t(\theta, z) = \mu + \sigma z$  with  $z \sim \mathcal{N}(0, \mathbf{I})$ .

246 **Bayesian ResNet-18 [He et al., 2016] on CIFAR-10 [Krizhevsky et al., 2012]:** We train here the  
 247 Bayesian variant of the ResNet-18 neural network introduced in [He et al., 2016] on CIFAR-10. The  
 248 latter dataset is composed of  $n = 60\,000$  handwritten digits,  $32 \times 32$  colour images in 10 classes,  
 249 with 6 000 images per class. As in the previous example, the weights are assumed independent and  
 250 identically distributed according to  $\mathcal{N}(0, \mathbf{I})$ . Standard hyperparameters values found in the literature,  
 251 such as the annealing constant or the number of MC samples, were used for the benchmark methods.  
 252 For better efficiency and lower variance, the Flipout estimator [Wen et al., 2018] is used.

253 **Experiment Results:** We compare the convergence of the *Monte Carlo variants* of the follow-  
 254 ing state of the art optimization algorithms — the ADAM [Kingma and Ba, 2015], the Momen-  
 255 tum [Sutskever et al., 2013] and the SAG [Schmidt et al., 2017] methods versus the *Bayes by Back-*  
 256 *prop* (BBB) [Blundell et al., 2015] and our proposed MISSO method. For all these methods, the  
 257 loss function (10) and its gradients were computed by Monte Carlo integration based on the re-  
 258 parametrization described above. The mini-batch of indices and MC samples are respectively set to  
 259 128 and  $M_{(k)} = k$ . The learning rates are set to  $10^{-3}$  for LeNet-5 and  $10^{-4}$  for Resnet-18.

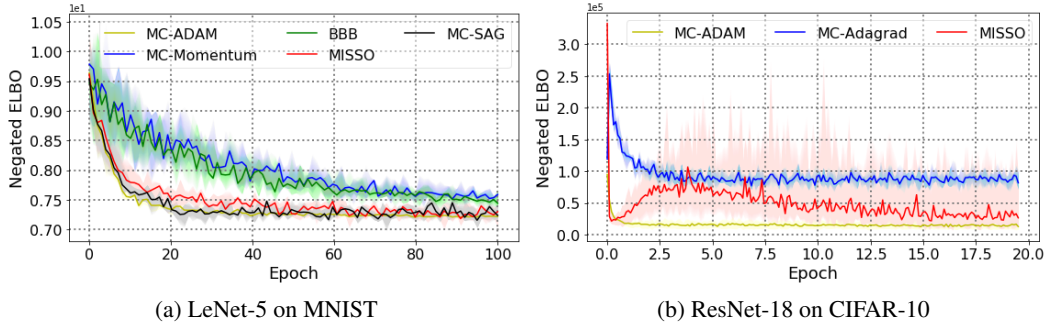


Figure 2: Negated ELBO versus epochs elapsed for fitting (a) Bayesian LeNet-5 on MNIST and (b) Bayesian ResNet-18 on CIFAR-10. The solid curve is obtained from averaging over 5 independent runs of the methods, and the shaded area represents the standard deviation.

260 Figure 2(a) shows the convergence of the negated evidence lower bound against the number of passes  
 261 over data (one pass represents an epoch). As observed, the proposed MISSO method outperforms  
 262 *Bayes by Backprop* and Momentum, while similar convergence rates are observed with the MISSO,  
 263 ADAM and SAG methods for our experiment on MNIST dataset using a Bayesian variant of LeNet-  
 264 5. On the other hand, the experiment conducted on CIFAR-10 (Figure 2(b)) using a much larger  
 265 network, *i.e.*, a Bayesian variant of ResNet-18 showcases the need of a well-tuned adaptive methods  
 266 to reach better training loss (and also faster). Our MISSO method is similar to the Monte Carlo  
 267 variant of ADAM but slower than Adagrad optimizer. Recall that the purpose of this paper is to  
 268 provide a common class of optimizers, such as VI, in order to study their convergence behaviors,  
 269 and not to introduce a novel method outperforming the baselines methods.

## 270 5 Conclusion

271 We present a unifying framework for minimizing a nonconvex and nonsmooth finite-sum objective  
 272 function using incremental surrogates when the latter functions are expressed as an expectation and  
 273 are intractable. Our approach covers a large class of nonconvex applications in machine learning  
 274 such as logistic regression with missing values and variational inference. We provide both finite-  
 275 time and asymptotic guarantees of our incremental stochastic surrogate optimization technique and  
 276 illustrate our findings training a binary logistic regression with missing covariates to predict hemor-  
 277 rhagic shock and Bayesian variants of two Convolutional Neural Networks on benchmark datasets.



## 6 Broader Impact Statement

The optimization problem tackled by this paper is rather general and includes many sensible and important use cases. The one we would like to focus on in this broader impact statement paragraph is the example on the TraumaBase dataset we develop in this paper, Section 4.1 for the prediction of hemorrhagic shocks. One issue, among others (like privacy and security of the data), when working with data coming from different sources (here different hospitals), appears to be the presence of missing values in each hospital’s database. To cope with that important challenge, we develop, in this paper, an algorithm that first simulates those missing values, during the computation of the surrogates, and then uses the complete data, comprised of the observations and the latter simulations, to update the model parameters iteratively. Of course, we would like to draw the attention on the various strategies to impute those missing values, each of them coming with their own advantages and ethical implications. Indeed, missing data pose an obvious threat to the quality of a study, for instance, when the missing data handling technique is prone to bias. We refer the readers to [Enders and Gottschall, 2011] for a comprehensive study of this research question.

## References

- 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>.
- C. K. Enders and A. C. Gottschall. The impact of missing data on the ethical quality of a research study. 2011.
- 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.

- 325 K. Lange. *MM Optimization Algorithms*. SIAM-Society for Industrial and Applied Mathematics,  
326 USA, 2016. ISBN 1611974399, 9781611974393.
- 327 Y. LeCun. The mnist database of handwritten digits. <http://yann.lecun.com/exdb/mnist/>, 1998.
- 328 Y. LeCun, L. Bottou, Y. Bengio, P. Haffner, et al. Gradient-based learning applied to document  
329 recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.
- 330 Y. Li and Y. Gal. Dropout inference in bayesian neural networks with alpha-divergences. In *Proceed-*  
331 *ings of the 34th International Conference on Machine Learning-Volume 70*, pages 2052–2061.  
332 JMLR. org, 2017.
- 333 J. Mairal. Incremental majorization-minimization optimization with application to large-scale ma-  
334 chine learning. *SIAM J. Optim.*, 25(2):829–855, 2015. ISSN 1052-6234. doi: 10.1137/  
335 140957639. URL <https://doi.org/10.1137/140957639>.
- 336 G. J. McLachlan and T. Krishnan. *The EM algorithm and extensions*. Wiley Series in Probabil-  
337 ity and Statistics. Wiley-Interscience [John Wiley & Sons], Hoboken, NJ, second edition, 2008.  
338 ISBN 978-0-471-20170-0. doi: 10.1002/9780470191613. URL [https://doi.org/10.1002/](https://doi.org/10.1002/9780470191613)  
339 [9780470191613](https://doi.org/10.1002/9780470191613).
- 340 S. P. Meyn and R. L. Tweedie. *Markov chains and stochastic stability*. Springer Science & Business  
341 Media, 2012.
- 342 R. M. Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business  
343 Media, 2012.
- 344 J. Paisley, D. Blei, and M. Jordan. Variational bayesian inference with stochastic search. In *ICML*.  
345 icml.cc / Omnipress, 2012.
- 346 N. G. Polson, V. Sokolov, et al. Deep learning: a bayesian perspective. *Bayesian Analysis*, 12(4):  
347 1275–1304, 2017.
- 348 X. Qian, A. Sailanbayev, K. Mishchenko, and P. Richtárik. Miso is making a comeback with better  
349 proofs and rates. *arXiv preprint arXiv:1906.01474*, 2019.
- 350 M. Razaviyayn, M. Hong, and Z.-Q. Luo. A unified convergence analysis of block successive  
351 minimization methods for nonsmooth optimization. *SIAM Journal on Optimization*, 23(2):1126–  
352 1153, 2013.
- 353 D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic backpropagation and approximate in-  
354 ference in deep generative models. In *International Conference on Machine Learning*, pages  
355 1278–1286, 2014.
- 356 M. Schmidt, N. Le Roux, and F. Bach. Minimizing finite sums with the stochastic average gradient.  
357 *Mathematical Programming*, 162(1-2):83–112, 2017.
- 358 I. Sutskever, J. Martens, G. Dahl, and G. Hinton. On the importance of initialization and momentum  
359 in deep learning. In *International conference on machine learning*, pages 1139–1147, 2013.
- 360 A. W. Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.
- 361 R. Vershynin. *High-dimensional probability: An introduction with applications in data science*,  
362 volume 47. Cambridge university press, 2018.
- 363 M. J. Wainwright. *High-dimensional statistics: A non-asymptotic viewpoint*, volume 48. Cambridge  
364 University Press, 2019.
- 365 G. C. G. Wei and M. A. Tanner. A monte carlo implementation of the em algorithm and the poor  
366 man’s data augmentation algorithms. *Journal of the American Statistical Association*, 85(411):  
367 699–704, 1990. doi: 10.1080/01621459.1990.10474930. URL [https://www.tandfonline.](https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474930)  
368 [com/doi/abs/10.1080/01621459.1990.10474930](https://www.tandfonline.com/doi/abs/10.1080/01621459.1990.10474930).
- 369 Y. Wen, P. Vicol, J. Ba, D. Tran, and R. Grosse. Flipout: Efficient pseudo-independent weight  
370 perturbations on mini-batches. *arXiv preprint arXiv:1803.04386*, 2018.