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:

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

where Θ is a convex, compact, and closed subset of $\mathbb{R}^p$, and for any $i \in [\![1, n]\!]$, the function $\mathcal{L}_i : \mathbb{R}^p \to \mathbb{R}$ is bounded from below and is (possibly) 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]. 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 new 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. Our contributions can be summarized as follows.

- 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 done under this unifying framework of analysis.
- We establish both asymptotic and non-asymptotic convergence for the MISSO method. In particular, the MISSO method converges almost surely to a stationary point and in $\mathcal{O}(n/\epsilon)$ iterations to an ϵ-stationary point.

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

Notations. We denote $[\![1,n]\!] = \{1,\ldots,n\}$. 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 $f : \Theta \to \mathbb{R}$, $f'(\boldsymbol{\theta}, \boldsymbol{d})$ is the directional derivative of f at $\boldsymbol{\theta}$ along the direction $\boldsymbol{d}$, *i.e.,*

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

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

2 Incremental Minimization of Finite Sum 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. The MM method tackles (1) through alternating between two steps — (i) minimizing a *surrogate* function which upper bounds the original objective function; and (ii) updating the surrogate function to tighten the upper bound.

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

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

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

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

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

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

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

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

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

Algorithm 1 MISO method [Mairal, 2015]

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

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

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

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

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

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

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

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

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

and we summarize the proposed MISSO method in Algorithm 2. As seen, the procedure is similar to the MISO method but it 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 $\widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta})$ used in lieu of $\mathcal{A}_i^k(\boldsymbol{\theta})$ in Line 6 when optimizing for the next iterate $\boldsymbol{\theta}^{(k)}$. We now discuss two applications of the MISSO method.

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

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

Algorithm 2 MISSO method

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

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

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

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

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

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

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

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) = \int_{\mathsf{Z}} \underbrace{\log \left(p_i(z_i, \overline{\boldsymbol{\theta}}) / f_i(z_i, \boldsymbol{\theta}) \right)}_{= r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i)} p_i(z_i, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_i) . \quad (7)$$

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

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

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

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

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

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

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

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

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \mathcal{L}_i(\overline{\boldsymbol{\theta}}) + \left\langle \nabla_{\boldsymbol{\theta}} \mathcal{L}_i(\overline{\boldsymbol{\theta}}) \,|\, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}} \right\rangle + \frac{\mathrm{L}}{2} \| \overline{\boldsymbol{\theta}} - \boldsymbol{\theta} \|^2, \ (\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) \in \Theta^2 . \quad (11)$$

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

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

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

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

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

3 Convergence Analysis

We now provide asymptotic and non-asymptotic convergence results for the MISSO method

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

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

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

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

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

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

Some intuitions behind the controlling terms: It is actually common in statistical and optimization problems, to deal with the manipulation and the control of random variables indexed by sets with an infinite number of elements. Here, the controlled random variable is an image of a continuous function defined as $r_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}},z_{i,m})-\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}})$ for all $z\in\mathsf{Z}$ and for fixed $(\boldsymbol{\theta},\overline{\boldsymbol{\theta}})\in\Theta^2$. 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 H2, which is classical in empirical processes. In [Vershynin, 2018, Theorem 8.2.3], the authors recall the uniform law of large numbers for a class of L-Lipschitz functions:

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

Moreover, in [Vershynin, 2018, Theorem 8.1.3] and [Wainwright, 2019, Theorem 5.22], the application of the Dudley's inequality yields:

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

where $\mathcal{N}(\mathcal{F},\|\cdot\|_\infty,\varepsilon)$ is the bracketing number and ϵ denotes the level of approximation (the bracketing number goes to infinity when $\epsilon\to0$). Finally, in [Van der Vaart, 2000, p.271, Example],

$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$ is bounded from above for a class of parametric functions $\mathcal{F} = f_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \Theta$ on a bounded set $\Theta \subset \mathbb{R}$:

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

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:

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

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

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

We first establish a non-asymptotic convergence rate for the MISSO method using a random termination number K as in [Ghadimi and Lan, 2013]:

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

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

Then we have following non-asymptotic bounds:

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

Note that $\Delta_{(K_{\mathsf{max}})}$ is finite for any $K_{\mathsf{max}} \in \mathbb{N}$. As expected, the MISSO method converges to a stationary point of (1) asymptotically and at a sublinear rate $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$. Furthermore,we remark that the MISO method can be analyzed in Theorem 1 as a special case of the MISSO method satisfying $C_{\mathsf{r}} = C_{\mathsf{gr}} = 0$. In this case, while the asymptotic convergence is well known from [Mairal, 2015] [cf. H2], Eq. (16) gives a non-asymptotic rate of $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{nL/K_{\mathsf{max}}})$ which is new to our best knowledge. Next, we show that under an additional assumption on the sequence of batch size $M_{(k)}$, the MISSO method converges almost surely to a stationary point:

Theorem 2. *Under S1, S2, H1, H2. In addition, assume that $\{M_{(k)}\}_{k \geq 0}$ 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., $\lim_{k \to \infty} g_-(\boldsymbol{\theta}^{(k)}) \overset{a.s.}{=} 0$.
2. *the objective value $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges a.s. to a finite number $\underline{\mathcal{L}}$,* i.e., $\lim_{k \to \infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \overset{a.s.}{=} \underline{\mathcal{L}}$.

In particular, the first result above shows that the sequence $\{\boldsymbol{\theta}^{(k)}\}_{k \geq 0}$ 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, $((y_i, z_i), i \in [\![n]\!])$ where $y_i \in \{0, 1\}$ is a binary response and $z_i = (z_{i,j} \in \mathbb{R}, j \in [\![p]\!])$ is a covariate vector. The vector of covariates $z_i = [z_{i,\text{mis}}, z_{i,\text{obs}}]$ 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 $(z_i, i \in [\![n]\!])$ are i.i.d. and marginally distributed according to $\mathcal{N}(\boldsymbol{\beta}, \boldsymbol{\Omega})$ where $\beta \in \mathbb{R}^p$ and Ω is a positive 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:

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

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

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

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. Extended implementation details are given in Appendix D.1.3.

Similar to [Jiang et al., 2018], we select $p = 16$ influential quantitative measurements, described in Appendix D.1.1, on $n = 6384$ patients, and we adopt the logistic regression model with missing covariates in (17) to predict the risk of a severe hemorrhage which is one of the main cause of death after a major trauma. For the Monte-Carlo sampling of $z_{i,\text{mis}}$, required while running MISSO, we run a Metropolis Hastings algorithm with the target distribution $p(\cdot|z_{i,\text{obs}}, y_i; \boldsymbol{\theta}^{(k)})$.

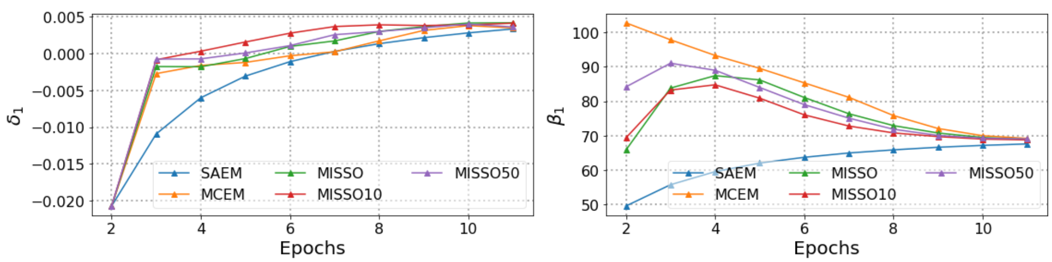


Figure 1: Convergence of first component of the vector of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against No. of passes over the data.

We compare in Figure 1 the convergence behavior of the estimated parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ using SAEM [Delyon et al., 1999] (with stepsize $\gamma_k = 1/k$), MCEM [Wei and Tanner, 1990] and the proposed MISSO method. For the MISSO method, we set the batch size to $M_{(k)} = 10 + k^2$ and we examine with selecting different number of functions in Line 5 in the method – the default settings with 1 (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 $[\![n]\!]$; step (ii) sample a Monte Carlo batch $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ from $\mathcal{N}(0, \mathbf{I})$; and step (iii) update the parameters, with $\tilde{w} = t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})$, as

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

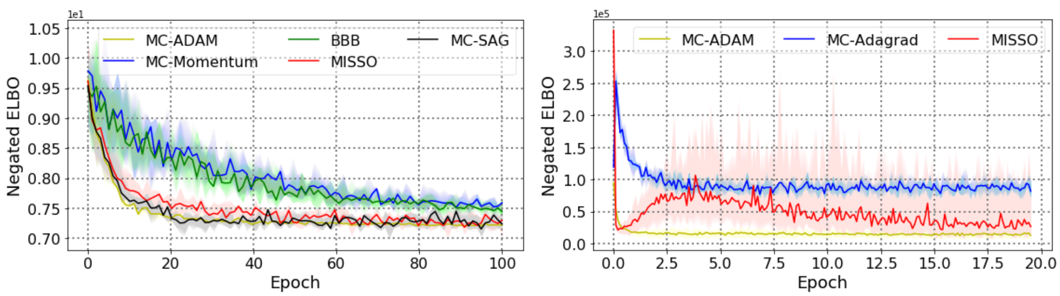
where $\hat{\mu}_\ell^{(\tau^k)} = \frac{1}{n}\sum_{i=1}^{n} \mu_\ell^{(\tau_i^k)}$ and $d(\boldsymbol{\theta}) = n^{-1}\sum_{\ell=1}^{d}\left(-\log(\sigma) + (\sigma^2 + \mu_\ell^2)/2 - 1/2\right)$.

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] (see Appendix D.2.1). We train this network on the MNIST dataset [LeCun, 1998]. The training set is composed of $n = 55\,000$ handwritten digits, 28×28

images. Each image is labelled with its corresponding number (from zero to nine). Under the prior distribution π, see (8), the weights are assumed independent and identically distributed according to $\mathcal{N}(0,1)$. We also assume that $q(\cdot;\boldsymbol{\theta}) \equiv \mathcal{N}(\mu, \sigma^2 \mathbf{I})$. The variational posterior parameters are thus $\boldsymbol{\theta} = (\mu, \sigma)$ where $\mu = (\mu_\ell, \ell \in \llbracket d \rrbracket)$ where d is the number of weights in the neural network. We use the re-parametrization as $w = t(\boldsymbol{\theta}, z) = \mu + \sigma z$ with $z \sim \mathcal{N}(0, \mathbf{I})$.

Bayesian ResNet-18 [He et al., 2016] on CIFAR-10 [Krizhevsky et al., 2012]: We train here the Bayesian variant of the ResNet-18 neural network (see Appendix D.2.2) introduced in [He et al., 2016] on CIFAR-10. The latter dataset is composed of $n = 60\,000$ handwritten digits, 32×32 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 $\mathcal{N}(0, \mathbf{I})$. The source code used as a backbone here can be found in the TensorFlow Probability Github repo where the default hyperparameters, such as the annealing constant or the number of MC samples, were used for the benchmark methods. For better efficiency and lower variance, the Flipout estimator [Wen et al., 2018] is 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 Backprop* (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 using Tensorflow Probability library [Dillon et al., 2017], based on the re-parametrization described above. Update rules for each algorithm are performed using their vanilla implementations on TensorFlow [Abadi et al., 2015] as detailed in Appendix D.2.3. We use the following hyperparameters for all runs — the learning rate is 10^{-3}, we run 100 epochs with a mini-batch size of 128 and use the batchsize of $M_{(k)} = k$.



(a) LeNet-5 on MNIST

(b) ResNet-18 on CIFAR-10

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.

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

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

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

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

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

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

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

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

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

S. Ghadimi and G. Lan. Stochastic first-and zeroth-order methods for nonconvex stochastic programming. *SIAM Journal on Optimization*, 23(4):2341–2368, 2013.

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

K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

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

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

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

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

A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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