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

Anonymous Author(s)
Affiliation
Address
email

Abstract

To be completed

1 Introduction

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

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

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

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 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 MISSO Algorithm

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*

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

Submitted to 34th Conference on Neural Information Processing Systems (NeurIPS 2020). Do not distribute.

Algorithm 1 MISSO method

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

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

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

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

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

Let Z be a measurable set, $p_i : \mathsf{Z} \times \Theta \to \mathbb{R}_+$ be a pdf, $r_i : \Theta \times \Theta \times \mathsf{Z} \to \mathbb{R}$ be a measurable function and μ_i be a σ-finite measure, we consider surrogate functions which satisfy S1, S2 that can be expressed as an expectation:

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

The MISSO method replaces the expectation in (5) by *Monte Carlo* integration and then optimizes (1) incrementally.

Denote by $M \in \mathbb{N}$ the Monte Carlo batch size and let $z_m \in \mathsf{Z}$, $m = 1, ..., M$ be a set of samples.

To this end, we define

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

and we summarize the proposed MISSO method in Algorithm 1.

3 Convergence Analysis

We provide non-asymptotic convergence bound 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.*

H2. *For all $i \in [\![1, n]\!]$, $(\theta, \overline{\boldsymbol{\theta}}) \in \Theta^2$, $z_i \in \mathsf{Z}$ we assume the existence of a majorizing function $m_{\mathsf{r}} : \mathsf{Z} \to \mathbb{R}$ and a constant $C_{\mathsf{r}} < \infty$ such that:*

$$\sup_{M>0} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \left\{ r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_{i,m}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \right\} < m_{\mathsf{r}}(z_i) \quad \textit{and} \quad \mathbb{E}_{\overline{\boldsymbol{\theta}}}\big[m_{\mathsf{r}}(z_i) | \mathcal{F}\big] < C_{\mathsf{r}} \tag{9}$$

where $\mathcal{F}$ is the filtration of the total randomness 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 $p_i(\cdot; \overline{\boldsymbol{\theta}})$. Besides,*

$$\sup_{M>0} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \left\{ \frac{\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}) - r_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}, z_{i,m})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \right\} < m_{\mathsf{gr}}(z_i) \quad \textit{and} \quad \mathbb{E}_{\overline{\boldsymbol{\theta}}}\big[m_{\mathsf{gr}}(z_i) | \mathcal{F}\big] < C_{\mathsf{gr}} \tag{10}$$

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

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

$$\mathbb{E} \sup_{f \in \mathcal{F}} \left| \frac{1}{M} \sum_{i=1}^{M} f(X_i) - \mathbb{E} f(X) \right| \leq \frac{CL}{\sqrt{M}} \tag{11}$$

Moreover, in [Vershynin, 2018] and [Wainwright, 2019], the application of the Dudley's inequality yields: $N_{11}\left(\varepsilon \|m\|_{P,r}, \mathcal{F}, L_r(P)\right) \leq K \left(\frac{\operatorname{diam} \Theta}{\varepsilon}\right)^d, \quad$ every $0 < \varepsilon < \operatorname{diam} \Theta$

$$\mathbb{E} \sup_{\substack{f \\ f \in \mathcal{F}}} |X_f| = \mathrm{E} \sup_{f \in \mathcal{F}} |X_f - X_0| \leq \frac{1}{\sqrt{M}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)} d\varepsilon \tag{12}$$

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 $\epsilon \to 0$). Finally, in [Van der Vaart, 2000], this bracketing number is upperbounded for a class of parametric function $\mathcal{F} = f_\theta : \theta \in \Theta$ on a bounded set $\Theta \subset \mathbb{R}$ as:

$$\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon) \leq K \left(\frac{\operatorname{diam} \Theta}{\varepsilon}\right)^d, \quad \text{every} \quad 0 < \varepsilon < \operatorname{diam} \Theta \tag{13}$$

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

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

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

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

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{15}$$

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

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

$$\Delta_{(K_{\mathsf{max}})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})] + \sum_{k=0}^{K_{\mathsf{max}}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}}, \tag{16}$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}\left[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\right] \leq \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}, \quad \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}. \tag{17}$$

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 $\mathbb{E}[g_{-}^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$.

Proof We begin by recalling the definition

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta}). \tag{18}$$

Notice that

$$\begin{aligned} \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) &= \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^{k+1})}, \{z_{i,m}^{(\tau_i^{k+1})}\}_{m=1}^{M_{(\tau_i^{k+1})}}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big). \end{aligned} \tag{19}$$

Furthermore, we recall that

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

Due to S2, we have

$$\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{21}$$

To prove the first bound in (17), using the optimality of $\boldsymbol{\theta}^{(k+1)}$, one has

$$\begin{aligned} &\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big) \end{aligned} \tag{22}$$

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

$$\begin{aligned} &\mathbb{E}_{i_k}\big[\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}})|\mathcal{F}_k, i_k\big]|\mathcal{F}_k\big] \\ &= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathbb{E}_{i_k}\big[\mathbb{E}\big[\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} r_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, z_{i_k,m}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})|\mathcal{F}_k, i_k\big]|\mathcal{F}_k\big] \\ &\leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}}, \end{aligned} \tag{23}$$

where the last inequality is due to H2. Moreover,

$$\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})|\mathcal{F}_k\big] = \frac{1}{n}\sum_{i=1}^{n} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, \{z_{i,m}^{(\tau_i^k)}\}_{m=1}^{M_{(\tau_i^k)}}) = \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{24}$$

Taking the conditional expectations on both sides of (22) and re-arranging terms give:

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} \tag{25}$$

Proceeding from (25), we observe the following lower bound for the left hand side

$$\begin{aligned} &\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \overset{(a)}{=} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \\ &\overset{(b)}{\geq} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2L}\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\ &= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} r_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})\Big\}}_{:= -\delta^{(k)}(\boldsymbol{\theta}^{(k)})} + \frac{1}{2L}\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \end{aligned} \tag{26}$$

where (a) is due to $\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$ [cf. S1], (b) is due to (21) and we have defined the summation in the last equality as $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$. Substituting the above into (25) yields

$$\frac{\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \le n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)}) \tag{27}$$

Observe the following upper bound on the total expectations:

$$\mathbb{E}\big[\delta^{(k)}(\boldsymbol{\theta}^{(k)})\big] \le \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} \frac{C_{\mathsf{r}}}{\sqrt{M_{(\tau_i^k)}}}\Big], \tag{28}$$

which is due to H2. It yields

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

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

$$\begin{aligned}\mathbb{E}\big[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] &= \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\ &\le \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{2LC_{\mathsf{r}}}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}\Big[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{M_{(\tau_i^k)}}}\Big]\end{aligned} \tag{29}$$

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

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^{k}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} \tag{30}$$

Taking the sum yields:

$$\begin{aligned}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] &= \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{j=1}^{k}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{l=0}^{k-1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{k-(l+1)} M_{(l)}^{-1/2} \\ &= \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}\sum_{k=l+1}^{K_{\mathsf{max}}-1}\frac{1}{n}\left(1 - \frac{1}{n}\right)^{k-(l+1)} \le \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}\end{aligned} \tag{31}$$

where the last inequality is due to upper bounding the geometric series. Plugging this back into (29) yields

$$\begin{aligned}\mathbb{E}\big[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] &= \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\ &\le \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}} = \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}.\end{aligned} \tag{32}$$

This concludes our proof for the first inequality in (17).

To prove the second inequality of (17), we define the shorthand notations $g^{(k)} := g(\boldsymbol{\theta}^{(k)})$, $g_-^{(k)} := -\min\{0, g^{(k)}\}$, $g_+^{(k)} := \max\{0, g^{(k)}\}$. We observe that

$$\begin{aligned} g^{(k)} &= \inf_{\boldsymbol{\theta}\in\Theta}\frac{\mathcal{L}'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \\ &= \inf_{\boldsymbol{\theta}\in\Theta}\Big\{\frac{\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} - \frac{\big\langle\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\,|\,\boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}\big\rangle}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|}\Big\} \\ &\ge -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta}\frac{\frac{1}{n}\sum_{i=1}^{n}\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|}\end{aligned} \tag{33}$$

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

$$
\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\
&= \underbrace{\widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})}_{\geq 0} - \widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}) + \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\
&\geq \frac{1}{n}\sum_{i=1}^{n} \left\{ \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) - \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) \right\}
\end{aligned}
\tag{34}
$$

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

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

We have

$$
g^{(k)} \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta} \in \Theta} (-\epsilon^{(k)}(\boldsymbol{\theta})) \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| - \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{35}
$$

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

$$
g_-^{(k)} \leq \|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{36}
$$

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

$$
\mathbb{E}[g_-^{(K)}] \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|] + \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \epsilon^{(K)}(\boldsymbol{\theta})] \tag{37}
$$

We note that

$$
\left(\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|]\right)^2 \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta(K_{\mathsf{max}})}{K_{\mathsf{max}}}, \tag{38}
$$

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

$$
\begin{aligned}
\mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \epsilon^{(K)}(\boldsymbol{\theta})] &= \frac{1}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}} \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \epsilon^{(k)}(\boldsymbol{\theta})] \overset{(a)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} M_{(\tau_i^k)}^{-1/2}\Big] \\
&\overset{(b)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}
\end{aligned}
\tag{39}
$$

where (a) is due to H2 and (b) is due to (31). This implies

$$
\mathbb{E}[g_-^{(K)}] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}, \tag{40}
$$

and concludes the proof of the theorem. □

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 $z_i = [z_{i,\mathrm{mis}}, z_{i,\mathrm{obs}}]$ is not fully observed where we denote by $z_{i,\mathrm{mis}}$

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:

$$p_i(y_i|z_i) = S(\boldsymbol{\delta}^\top \bar{z}_i)^{y_i} \left(1 - S(\boldsymbol{\delta}^\top \bar{z}_i)\right)^{1-y_i} \tag{41}$$

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 $\bar{z}_i = (1, z_i)$. We are interested in estimating $\boldsymbol{\delta}$ and finding the latent structure of the covariates z_i. Here, $\boldsymbol{\theta} = (\boldsymbol{\delta}, \boldsymbol{\beta}, \boldsymbol{\Omega})$ is the parameter to estimate. For $i \in [\![n]\!]$, the complete data log-likelihood is expressed as:

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

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

$$\begin{aligned}
&\boldsymbol{\beta}^{(k)} = \arg\min_{\beta \in \Theta} \frac{1}{n}\sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega^{(k)}, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(k)} \\
&\boldsymbol{\Omega}^{(k)} = \arg\min_{\Omega \in \Theta} \frac{1}{n}\sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta^{(k)}, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n}\sum_{i=1}^n \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(k)}(z_{i,m}^{(k)})^\top - \boldsymbol{\beta}^{(k)}(\boldsymbol{\beta}^{(k)})^\top \\
&\boldsymbol{\delta}^{(k)} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{\delta}^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1}\tilde{D}^{(k)}.
\end{aligned} \tag{42}$$

where $z_{i,m}^{(k)} = (z_{i,\text{mis},m}^{(k)}, z_{i,\text{obs}})$ is composed of a simulated and an observed part and $\tilde{D}^{(k)} = \frac{1}{n}\sum_{i=1}^n \tilde{D}_i^{(\tau_i^k)}$ and $\tilde{H}^{(k)} = \frac{1}{n}\sum_{i=1}^n \tilde{H}_i^{(\tau_i^k)}$. Besides, $\tilde{\mathcal{L}}_i^{(1)}(\beta, \Omega, \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M)$ and $\tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M)$ are defined as MC approximation of $\hat{\mathcal{L}}_i^{(1)}(\beta, \Omega, \overline{\boldsymbol{\theta}})$ and $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \overline{\boldsymbol{\theta}})$, for all $i \in [\![n]\!]$.

See Appendix **??** for more explanation.

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

Similar to [Jiang et al., 2018], we select $p = 16$ influential quantitative measurements, described in Appendix **??**, on $n = 6384$ patients, and we adopt the logistic regression model with missing covariates in (41) to predict the risk of a severe hemorrhage which is one of the main cause of death after a major trauma. Note as the dataset considered is heterogeneous – coming from multiple sources with frequently missed entries – we apply the latent data model described in the above. For the Monte-Carlo sampling of $z_{i,\text{mis}}$, we run a Metropolis Hastings algorithm with the target distribution $p(\cdot|z_{i,\text{obs}}, y_i; \boldsymbol{\theta}^{(k)})$ whose procedure is detailed in Appendix **??**.

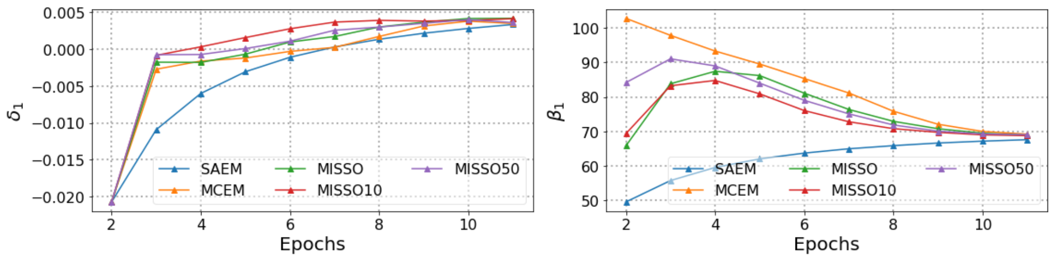


Figure 1: Convergence of first component of the vector of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against the number of passes over the data.

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

4.2 Training Bayesian CNN using MISSO

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

$$\mu_\ell^{(k)} = \frac{1}{n}\sum_{i=1}^n \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^n \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n}\sum_{i=1}^n \sigma^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^n \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} \,, \tag{43}$$

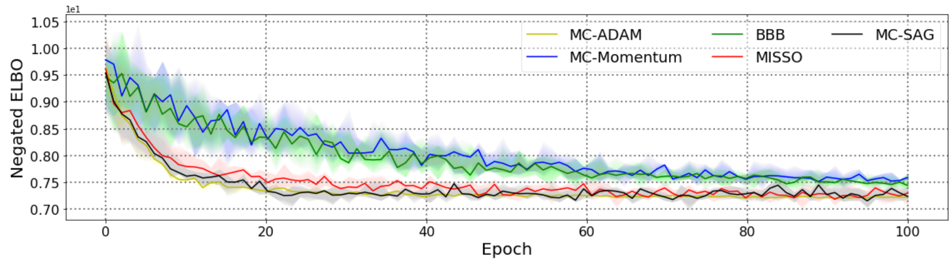
where $\hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} = \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k-1)}$ and $\hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} = \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k-1)}$ for $i \neq i_k$ and:

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

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

Bayesian LeNet-5 on MNIST: This application follows **Example 2** described in Section 2. We apply the MISSO method to fit a Bayesian variant of LeNet-5 [LeCun et al., 1998] (see Appendix **??**). 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 (**??**), the weights are assumed independent and identically distributed according to $\mathcal{N}(0, 1)$. We also assume that $q(\cdot; \boldsymbol{\theta}) \equiv \mathcal{N}(\mu, \sigma^2 \mathbf{I})$. The variational posterior parameters are thus $\boldsymbol{\theta} = (\mu, \sigma)$ where $\mu = (\mu_\ell, \ell \in [\![d]\!])$ where d is the number of weights in the neural network. We use the re-parametrization as $w = t(\boldsymbol{\theta}, z) = \mu + \sigma z$ with $z \sim \mathcal{N}(0, \mathbf{I})$.

We describe in Table 4.2 the architecture of the Convolutional Neural Network introduced in [LeCun et al., 1998] and trained on MNIST:



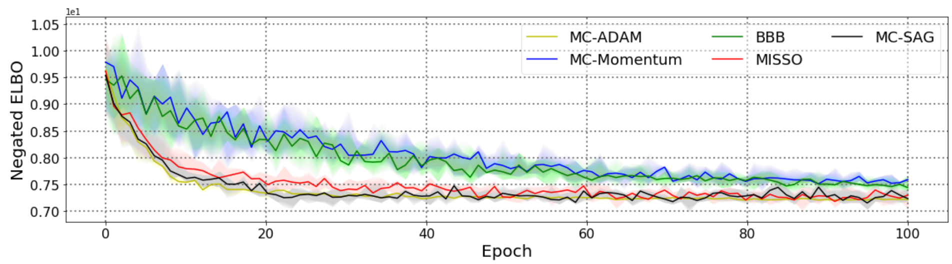


Figure 2: (Top) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. (Bottom) ELBO versus epochs elapsed for fitting the Bayesian ResNet-18 on CIFAR-10 using different algorithms. The solid curve is obtained from averaging over 5 independent runs of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

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

Table 1: LeNet-5 architecture

Bayesian ResNet-18 on CIFAR-10: Put here the architectureof ResNet-18 introduced in [He et al., 2016].

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 (**??**) 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 **??**. We use the following hyperparameters for all runs — the learning rate is 10^{-3}, we run 100 epochs with a mini-batch size of 128 and use the batchsize of $M_{(k)} = k$. Figure 2 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.

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. 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.

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.

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`.

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.

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`.