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

A Proofs of the Theoretical Results

A.1 Proof of Theorem 1

Theorem. *Under H1-H4. 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:

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

Proof We begin by recalling the definition

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

Notice that

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

Furthermore, we recall that

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

Due to H2, we have

$$\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \ . \tag{19}$$

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

$$\begin{aligned}&\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big) \ .\end{aligned} \tag{20}$$

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

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

where the last inequality is due to H4. Moreover,

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

Taking the conditional expectations on both sides of (20) and re-arranging terms give:

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} \ . \tag{21}$$

Proceeding from (21), we observe the following lower bound for the left hand side

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

where (a) is due to $\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$ [cf. H1], (b) is due to (19) and we have defined the summation in the last equality as $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$. Substituting the above into (21) yields

$$\frac{\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)}) \ . \tag{22}$$

Observe the following upper bound on the total expectations:

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

which is due to H4. It yields

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

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 H4 and taking total expectations lead to

$$\begin{aligned}
&\mathbb{E}\big[\|\nabla\widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] = \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{2LC_{\mathsf{r}}}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}\Big[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{M_{(\tau_i^k)}}}\Big] \ .
\end{aligned} \tag{23}$$

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:

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^{k}\frac{1}{n}\left(1-\frac{1}{n}\right)^{j-1}M_{(k-j)}^{-1/2} \tag{24}$$

Taking the sum yields:

$$\begin{aligned}
&\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{j=1}^{k}\frac{1}{n}\left(1-\frac{1}{n}\right)^{j-1}M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{l=0}^{k-1}\frac{1}{n}\left(1-\frac{1}{n}\right)^{k-(l+1)}M_{(l)}^{-1/2} \\
&= \sum_{l=0}^{K_{\mathsf{max}}-1}M_{(l)}^{-1/2}\sum_{k=l+1}^{K_{\mathsf{max}}-1}\frac{1}{n}\left(1-\frac{1}{n}\right)^{k-(l+1)} \leq \sum_{l=0}^{K_{\mathsf{max}}-1}M_{(l)}^{-1/2} \ ,
\end{aligned} \tag{25}$$

where the last inequality is due to upper bounding the geometric series. Plugging this back into (23) yields

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

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

To prove the second inequality of (16), 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

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

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 $\boldsymbol{\theta}\in\Theta$,

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

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

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

We have

$$g^{(k)} \geq -\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta}(-\epsilon^{(k)}(\boldsymbol{\theta})) \geq -\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| - \sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|\ . \tag{26}$$

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

$$g_-^{(k)} \leq \|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})|\ . \tag{27}$$

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

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

We note that

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

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

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

where (a) is due to H4 and (b) is due to (25). This implies

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

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

A.2 Proof of Theorem 2

Theorem. *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., $\lim_{k\to\infty} g_-(\boldsymbol{\theta}^{(k)}) \overset{a.s.}{=} 0$.
2. *the objective value $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges a.s. to a finite number $\underline{\mathcal{L}}$,* i.e., $\lim_{k\to\infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \overset{a.s.}{=} \underline{\mathcal{L}}$.

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

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

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1} \tag{28}$$

then:

(i) for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.

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

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

We proceed from (20) by re-arranging terms and observing that

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

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

$$V_k := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \inf_{k\geq 0} c_k \geq 0 \tag{29}$$

is a non-negative random variable.

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

$$X_k := \tfrac{1}{n}\big(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(\tau_{i_k}^k)}; \boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})\big) \geq 0 \; . \tag{30}$$

Thirdly, we define

$$\begin{aligned}
E_k = &-\big(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})\big) + \big(\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\big) \\
&+ \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})\big) \\
&+ \tfrac{1}{n}\big(\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big) \; .
\end{aligned} \tag{31}$$

Note that from the definitions (29), (30), (31), we have $V_{k+1} \leq V_k - X_k + E_k$ for any $k \geq 1$.

Under H4, we observe that

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

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

$$\mathbb{E}\big[|\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})|\big] \leq \tfrac{1}{n}\textstyle\sum_{i=1}^n C_{\mathsf{r}}\mathbb{E}\Big[M_{(\tau_i^k)}^{-1/2}\Big]$$

Therefore,

$$\mathbb{E}\big[|E_k|\big] \leq \tfrac{C_{\mathsf{r}}}{n}\Big(M_{(k)}^{-1/2} + \mathbb{E}\Big[M_{(\tau_{i_k}^k)}^{-1/2} + \textstyle\sum_{i=1}^n\big\{M_{(\tau_i^k)}^{-1/2} + M_{(\tau_i^{k+1})}^{-1/2}\big\}\Big]\Big) \ .$$

Using (25) and the assumption on the sequence $\{M_{(k)}\}_{k\geq 0}$, we obtain that

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

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

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

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

$$\lim_{k\to\infty}\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0 \quad \text{a.s.} \tag{32}$$

and subsequently applying (19), we have $\lim_{k\to\infty}\|\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| = 0$ almost surely. Finally, it follows from (19) and (27) that

$$\lim_{k\to\infty} g_{-}^{(k)} \leq \lim_{k\to\infty}\sqrt{2L}\sqrt{\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})} + \lim_{k\to\infty}\sup_{\boldsymbol{\theta}\in\Theta}|\epsilon^{(k)}(\boldsymbol{\theta})| = 0 \ , \tag{33}$$

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

Finally, we prove that $\mathcal{L}(\boldsymbol{\theta}^{(k)})$ converges almost surely. As a consequence of Lemma 1, it is clear that $\{V_k\}_{k\geq 0}$ converges almost surely and so is $\{\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\}_{k\geq 0}$, *i.e.*, we have $\lim_{k\to\infty}\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$. Applying (32) implies that

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

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

A.3 Proof of Lemma 1

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

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

then:

(i) for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.

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

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

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

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \ , \tag{34}$$

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

Since $0 \leq X_k \leq V_{k-1} - V_k + E_k$ we also obtain for all $k \geq 0$, $\mathbb{E}[X_k] < \infty$. Moreover, since $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$, the series $\sum_{j=1}^{\infty} E_j$ converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \tag{35}$$

Note that $\mathbb{E}[|W_k|] \leq \mathbb{E}[V_k] + \mathbb{E}\left[\sum_{j=k+1}^{\infty} |E_j|\right] < \infty$. For all $k \geq 1$, we get:

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \ . \end{aligned} \tag{36}$$

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

$$\mathbb{E}[\liminf_{k\to\infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k\to\infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \ , \tag{37}$$

showing that the random variable W_∞ is integrable.

In the sequel, set $U_k \triangleq W_0 - W_k$. By construction we have for all $k \geq 0$, $U_k \geq 0$, $U_k \leq U_{k+1}$ and $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$ and by the monotone convergence theorem, we get:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k\to\infty} U_k] \ . \tag{38}$$

Finally, we have:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k\to\infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \ . \tag{39}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (36) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\begin{aligned} &\sum_{j=1}^{\infty} X_j \leq W_0 - W_\infty < \infty \ , \\ &\sum_{j=1}^{\infty} \mathbb{E}[X_j] \leq \mathbb{E}[W_0] - w_\infty < \infty \ , \end{aligned} \tag{40}$$

an concludes the proof of the lemma. □

B Practical Details for the Binary Logistic Regression on the Traumabase

B.1 Traumabase dataset quantitative variables

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

B.2 Metropolis-Hastings algorithm

During the simulation step of the MISSO method, the sampling from the target distribution $\pi(z_{i,\text{mis}};\boldsymbol{\theta}) := p(z_{i,\text{mis}}|z_{i,\text{obs}},y_i;\boldsymbol{\theta})$ is performed using a Metropolis-Hastings (MH) algorithm [19] with proposal distribution $q(z_{i,\text{mis}};\boldsymbol{\delta}) := p(z_{i,\text{mis}}|z_{i,\text{obs}};\boldsymbol{\delta})$ where $\boldsymbol{\theta} = (\beta, \Omega)$ and $\boldsymbol{\delta} = (\xi, \Sigma)$. The parameters of the Gaussian conditional distribution of $z_{i,\text{mis}}|z_{i,\text{obs}}$ read:

$$\xi = \beta_{miss} + \Omega_{mis,obs}\Omega^{-1}_{obs,obs}(z_{i,\text{obs}} - \beta_{obs}) \,,$$
$$\Sigma = \Omega_{mis,mis} + \Omega_{mis,obs}\Omega^{-1}_{obs,obs}\Omega_{obs,\text{mis}} \,,$$

where we have used the Schur Complement of $\Omega_{obs,obs}$ in Ω and noted β_{mis} (resp. β_{obs}) the missing (resp. observed) elements of β. The MH algorithm is summarized in Algorithm 3.

Algorithm 3 MH aglorithm

```
Input: initialization z_{i,mis,0} ~ q(z_{i,mis}; δ)
for m = 1, ··· , M do
    Sample z_{i,mis,m} ~ q(z_{i,mis}; δ)
    Sample u ~ U([[0, 1]])
    Calculate the ratio r = [π(z_{i,mis,m}; θ)/q(z_{i,mis,m}); δ)] / [π(z_{i,mis,m-1}; θ)/q(z_{i,mis,m-1}); δ)]
    if u < r then
        Accept z_{i,mis,m}
    else
        z_{i,mis,m} ← z_{i,mis,m-1}
    end if
end for
Output: z_{i,mis,M}
```

B.3 MISSO Update

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

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}) = \int_{\mathsf{Z}} \log\left(p_i(z_{i,\text{mis}},\overline{\boldsymbol{\theta}})/f_i(z_{i,\text{mis}},\boldsymbol{\theta})\right) p_i(z_{i,\text{mis}},\overline{\boldsymbol{\theta}})\mu_i(\mathrm{d}z_i) \,.$$

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

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

$$\begin{aligned}
&\widehat{\mathcal{L}}_i(\boldsymbol{\theta};\overline{\boldsymbol{\theta}}) \\
=&- \int_{\mathsf{Z}} \log f_i(z_{i,\text{mis}}, z_{i,\text{obs}},\boldsymbol{\theta}) p_i(z_{i,\text{mis}},\overline{\boldsymbol{\theta}})\mu_i(\mathrm{d}z_{i,\text{mis}}) \\
=&- \int_{\mathsf{Z}} \log\left[p_i(y_i|z_{i,\text{mis}}, z_{i,\text{obs}},\delta)p_i(z_{i,\text{mis}},\beta,\Omega)\right] p_i(z_i,\overline{\boldsymbol{\theta}})\mu_i(\mathrm{d}z_{i,\text{mis}}) \\
=&\underbrace{- \int_{\mathsf{Z}} \log p_i(y_i|z_{i,\text{mis}}, z_{i,\text{obs}},\delta) p_i(z_i,\overline{\boldsymbol{\theta}})\mu_i(\mathrm{d}z_{i,\text{mis}})}_{=\hat{\mathcal{L}}^{(1)}_i(\delta,\overline{\boldsymbol{\theta}})} \underbrace{- \int_{\mathsf{Z}} \log p_i(z_{i,\text{mis}},\beta,\Omega) p_i(z_i,\overline{\boldsymbol{\theta}})\mu_i(\mathrm{d}z_{i,\text{mis}})}_{=\hat{\mathcal{L}}^{(2)}_i(\beta,\Omega,\overline{\boldsymbol{\theta}})} \,.
\end{aligned} \tag{41}$$

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

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

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

Recall that:

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

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

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

Then, we use, for all $i \in [\![n]\!]$, the following surrogate function to estimate δ:

$$\bar{\mathcal{L}}_i^{(1)}(\delta, \overline{\boldsymbol{\theta}}) = \hat{\mathcal{L}}_i^{(1)}(\bar{\delta}, \overline{\boldsymbol{\theta}}) - D_i^\top(\delta - \bar{\delta}) + \frac{1}{2}(\delta - \bar{\delta}) H_i (\delta - \bar{\delta})^\top \ , \tag{42}$$

where:

$$D_i = \int_{\mathsf{Z}} \nabla \log p_i(y_i|z_{i,\text{mis}}, z_{i,\text{obs}}, \delta)\big|_{\delta=\bar{\delta}} p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}}) \ ,$$
$$H_i = \int_{\mathsf{Z}} H_i(z_{i,\text{mis}}) p_i(z_{i,\text{mis}}, \overline{\boldsymbol{\theta}}) \mu_i(\mathrm{d}z_{i,\text{mis}}) \ .$$

Finally, at iteration k, the total surrogate is:

$$\begin{aligned}
\tilde{\mathcal{L}}^{(k)}(\theta) &= \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathcal{L}}_i(\theta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) \\
&= \frac{1}{n} \sum_{i=1}^{n} \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) - \frac{1}{n} \sum_{i=1}^{n} \tilde{D}_i^{(\tau_i^k)}(\delta - \delta^{(\tau_i^k)}) \\
&+ \frac{1}{2n} \sum_{i=1}^{n} (\delta - \delta^{(\tau_i^k)}) \left\{ \tilde{H}_i^{(\tau_i^k)} \right\} (\delta - \delta^{(\tau_i^k)})^\top \ ,
\end{aligned} \tag{43}$$

where for all $i \in [\![n]\!]$:

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

Minimizing the total surrogate (43) boils down to performing a quasi-Newton step. It is perhaps sensible to apply some diagonal loading which is perfectly compatible with the surrogate interpretation we just gave.

The logistic parameters are estimated as follows:

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

where $\tilde{\mathcal{L}}_i^{(1)}(\delta, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}})$ is the MC approximation of the MISO surrogate defined in (42) and which leads to the following quasi-Newton step:

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

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

MISSO updates: 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

$\{z_{i_k,\mathrm{mis},m}^{(k)}\}_{m=1}^{M_{(k)}}$ of missing values from the conditional distribution $p(z_{i_k,\mathrm{mis}}|z_{i_k,\mathrm{obs}},y_{i_k};\boldsymbol{\theta}^{(k-1)})$ using an MCMC sampler and computing the estimated parameters as follows:

$$
\begin{aligned}
\boldsymbol{\beta}^{(k)} &= \arg\min_{\beta\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathcal{L}}_i^{(2)}(\beta,\Omega^{(k)},\theta^{(\tau_i^k)},\{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} z_{i,m}^{(k)}\,,\\
\boldsymbol{\Omega}^{(k)} &= \arg\min_{\Omega\in\Theta} \frac{1}{n}\sum_{i=1}^{n} \tilde{\mathcal{L}}_i^{(2)}(\beta^{(k)},\Omega,\theta^{(\tau_i^k)},\{z_{i,m}\}_{m=1}^{M_{(\tau_i^k)}}) = \frac{1}{n}\sum_{i=1}^{n}\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} w_{i,m}^{(k)}\,,\\
\boldsymbol{\delta}^{(k)} &= \frac{1}{n}\sum_{i=1}^{n}\boldsymbol{\delta}^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1}\tilde{D}^{(k)}\,.
\end{aligned} \tag{44}
$$

where $z_{i,m}^{(k)} = (z_{i,\mathrm{mis},m}^{(k)}, z_{i,\mathrm{obs}})$ is composed of a simulated and an observed part, $\tilde{D}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{D}_i^{(\tau_i^k)}$, $\tilde{H}^{(k)} = \frac{1}{n}\sum_{i=1}^{n}\tilde{H}_i^{(\tau_i^k)}$ and $w_{i,m}^{(k)} = z_{i,m}^{(k)}(z_{i,m}^{(k)})^\top - \boldsymbol{\beta}^{(k)}(\boldsymbol{\beta}^{(k)})^\top$. 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 $i\in[\![n]\!]$ as components of the surrogate function (41).

B.4 Wall clock time

We provide Table 1, the running time for each method, plotted in Figure 1, employed to train a logistic regression with missing values on the TraumaBase dataset ($p = 16$ influential quantitative measurements, on $n = 6384$ patients).

The running times are sensibly the same since for each method the computation complexity per epoch is similar. We remark a slight delay using the MISSO method with a batch size of 1, as our code implemented in R, is not totally optimized and parallelized. Yet, when the batch size tends to 100%, we retrieve the duration of MCEM, which is consistent with the fact that MISSO with a full batch update boils down to the MCEM algorithm.

Table 1: Logistic Regression with missing values: running time in seconds for 10 epochs.

	SAEM	MCEM	MISSO	MISSO10	MISSO50
Logistic Regression	2033.2	1972.4	2244.8	2139.4	2005.2

We plot Figure 4, the updated parameters for the Logistic regression example against the time elapsed (in seconds).

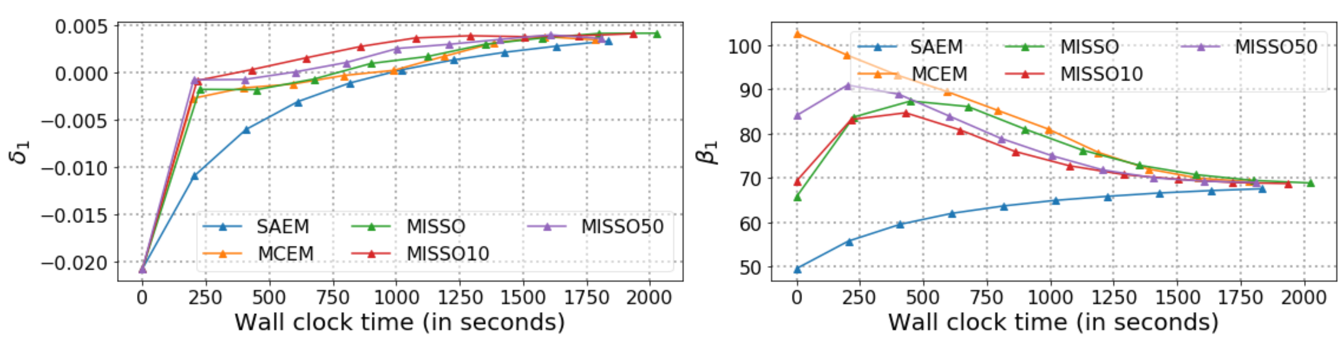


Figure 3: Convergence of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against time elapsed (in seconds).

B.5 Plots against epochs elapsed

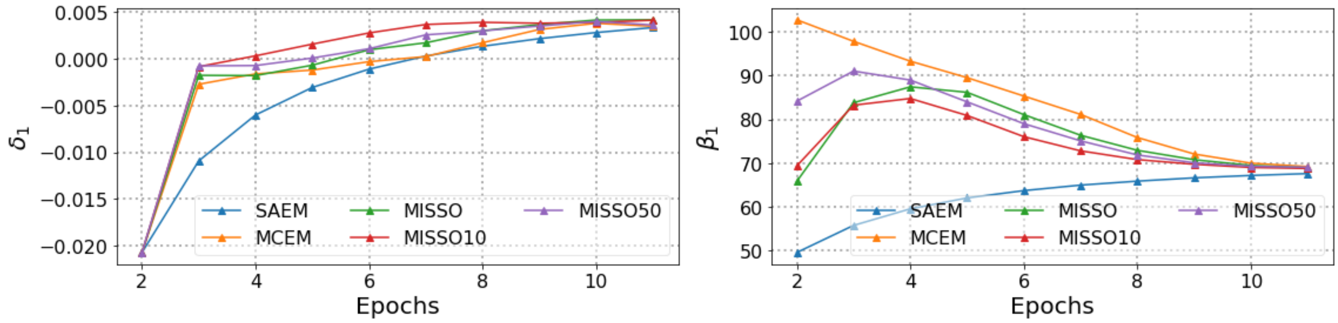


Figure 4: Convergence of parameters $\boldsymbol{\delta}$ and $\boldsymbol{\beta}$ for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against time elapsed (in seconds).

C Practical Details for the Incremental Variational Inference

C.1 Neural Networks Architecture

Bayesian LeNet-5 Architecture: We describe in Table 2 the architecture of the Convolutional Neural Network introduced in [15] and trained on MNIST:

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 2: LeNet-5 architecture

Bayesian ResNet-18 Architecture: We describe in Table 3 the architecture of the Resnet-18 we train on CIFAR-10:

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

Table 3: ResNet-18 architecture

C.2 Algorithms updates

First, we initialize the means $\mu_\ell^{(0)}$ for $\ell \in [\![d]\!]$ and variance estimates $\sigma^{(0)}$. 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 as

$$\mu_\ell^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n}\sum_{i=1}^{n} \sigma^{(\tau_i^k)} - \frac{\gamma}{n}\sum_{i=1}^{n} \hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} \; , \tag{45}$$

where we define the following gradient terms for all $i \in [\![1, n]\!]$:

$$\begin{aligned}
\hat{\boldsymbol{\delta}}_{\mu_\ell,i}^{(k)} &= -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} \nabla_w \log p(y_i|x_i, w)\Big|_{w=t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\mu_\ell} d(\boldsymbol{\theta}^{(k-1)}) \; , \\
\hat{\boldsymbol{\delta}}_{\sigma,i}^{(k)} &= -\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} z_m^{(k)} \nabla_w \log p(y_i|x_i, w)\Big|_{w=t(\boldsymbol{\theta}^{(k-1)}, z_m^{(k)})} + \nabla_{\sigma} d(\boldsymbol{\theta}^{(k-1)}) \; .
\end{aligned} \tag{46}$$

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

For all benchmark algorithms, we pick, at iteration k, a function index i_k uniformly on $[\![n]\!]$ and sample a Monte Carlo batch $\{z_m^{(k)}\}_{m=1}^{M_{(k)}}$ from the standard Gaussian distribution. The updates of the parameters μ_ℓ for all $\ell \in [\![d]\!]$ and σ break down as follows:

Monte Carlo SAG update: Set

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

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 are defined by (46) for $i = i_k$. The learning rate is set to $\gamma = 10^{-3}$.

Bayes By Backprop update: Set

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

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

Monte Carlo Momentum update: Set

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

where

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

where α and γ, respectively the momentum and the learning rates, are set to 10^{-3}.

Monte Carlo ADAM update: Set

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

where

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

and

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

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

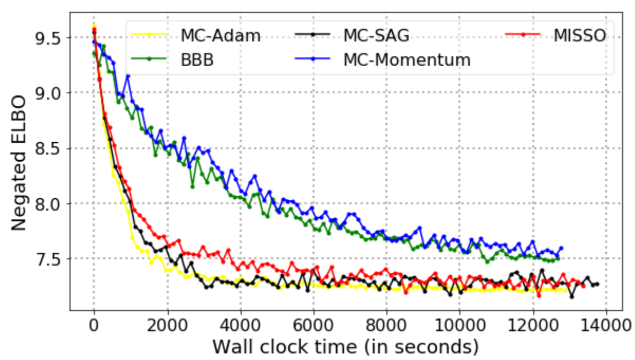
C.3 Wall clock time

LeNet-5 on MNIST We provide Table 4, the running time for each method, plotted in Figure 7, used to train a Bayesian variant of LeNet-5 on MNIST. The incremental method as MISSO and MC-SAG displays a similar wall clock time, despite being a bit worse given (a) the initialization that requires to compute a vector of n gradients kept in memory and updated through the iterations and (b) the average operation for each parameters update to compute the aggregated drift term, see (45).

Table 4: Bayesian Deep Neural Network: running time in seconds for 100 epochs.

	MC-Adam	MC-Momentum	BBB	MC-SAG	MISSO
LeNet-5 on MNIST	12889	12816	12690	13822	13367

We plot Figure 5, the learning curves for the MNIST example against the time elapsed (in seconds).



(a) LeNet-5 on MNIST

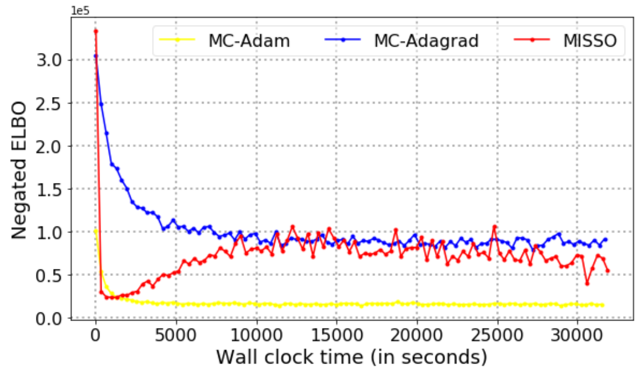
Figure 5: Negated ELBO versus wall clock time for fitting a Bayesian LeNet-5 on MNIST. Plotted on the average of the 5 repetitions.

ResNet-18 on CIFAR10 We provide Table 5, the running time for each method, plotted in Figure 7, used to train a Bayesian variant of ResNet-18 on CIFAR10.

Table 5: ResNet-18: running time in seconds for 20 epochs.

	MC-Adam	MC-Adagrad	MISSO
ResNet-18 on CIFAR	31879	31675	32092

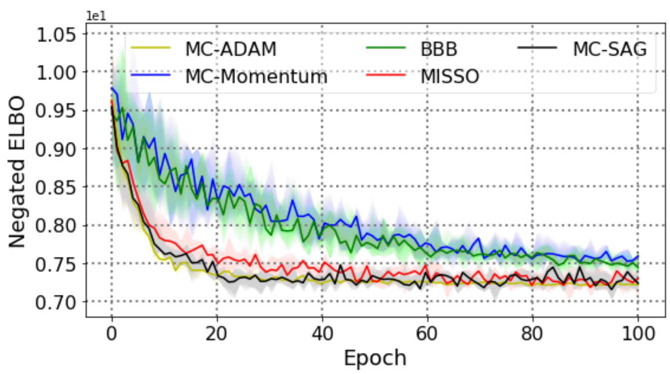
We plot Figure 6, the learning curves for the MNIST example against the time elapsed (in seconds).



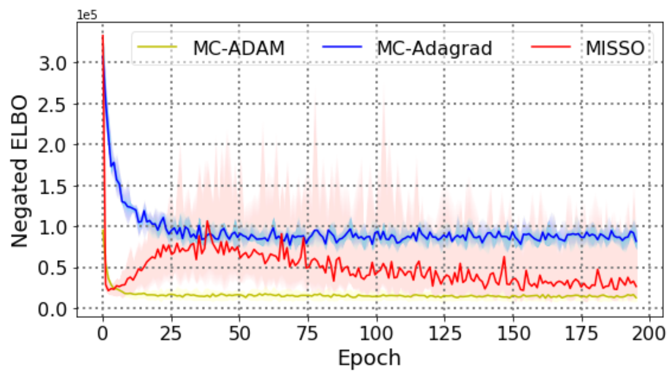
(a) ResNet-18 on CIFAR10

Figure 6: Negated ELBO versus wall clock time for fitting a Bayesian ResNet-18 on CIFAR10. Plotted on the average of the 5 repetitions.

C.4 Plots against the epochs elapsed



(a) LeNet-5: MNIST



(b) ResNet-18: CIFAR-10

Figure 7: 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.