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

$$C_{\mathsf{r}} := \sup_{\overline{\boldsymbol{\theta}} \in \Theta} \sup_{M > 0} \frac{1}{\sqrt{M}} \mathbb{E}_{\overline{\boldsymbol{\theta}}} \left[\sup_{\boldsymbol{\theta} \in \Theta} \left| \sum_{m=1}^{M} \left\{ r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_{i,m}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \right\} \right| \right] \tag{9}$$

$$C_{\mathsf{gr}} := \sup_{\overline{\boldsymbol{\theta}} \in \Theta} \sup_{M > 0} \sqrt{M} \mathbb{E}_{\overline{\boldsymbol{\theta}}} \left[\sup_{\boldsymbol{\theta} \in \Theta} \left| \frac{1}{M} \sum_{m=1}^{M} \frac{\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}) - r_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}, z_{i,m})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \right|^2 \right] \tag{10}$$

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 $p_i(\cdot; \overline{\boldsymbol{\theta}})$.*

H3. *For all $i \in [\![1, n]\!]$, $(\theta, \overline{\boldsymbol{\theta}}) \in \Theta^2$, $z_i \in \mathsf{Z}$ we assume the existence of a function $C_{\mathsf{r}} : \mathsf{Z} \to \mathbb{R}$ such that:*

$$\sup_{M>0} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \left\{ r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_{i,m}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \right\} < C_{\mathsf{r}}(z_i) \\ \mathbb{E}_{\overline{\boldsymbol{\theta}}}[C_{\mathsf{r}}(z_i)|\mathcal{F}] < \infty \tag{11}$$

where $\mathcal{F}$ is the filtration of the total randomness. Besides,

$$\sup_{M>0} \frac{1}{\sqrt{M}} \sum_{m=1}^{M} \left\{ \frac{\widehat{\mathcal{L}}_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}) - r_i'(\boldsymbol{\theta}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}}; \overline{\boldsymbol{\theta}}, z_{i,m})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \right\} < C_{\mathsf{gr}}(z_i) \\ \mathbb{E}_{\overline{\boldsymbol{\theta}}}[C_{\mathsf{gr}}(z_i)|\mathcal{F}] < \infty \tag{12}$$

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

$$g(\overline{\boldsymbol{\theta}}) := \inf_{\boldsymbol{\theta} \in \Theta} \frac{\mathcal{L}'(\overline{\boldsymbol{\theta}}, \boldsymbol{\theta} - \overline{\boldsymbol{\theta}})}{\|\overline{\boldsymbol{\theta}} - \boldsymbol{\theta}\|} \quad \text{and} \quad g(\overline{\boldsymbol{\theta}}) = g_+(\overline{\boldsymbol{\theta}}) - g_-(\overline{\boldsymbol{\theta}}) , \tag{13}$$

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 $g_-(\overline{\boldsymbol{\theta}}) = 0$ [**?**].

Also, denote

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{14}$$

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

Theorem 1. *Under S1, S2, H1, H3. 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:*

$$\Delta_{(K_{\mathsf{max}})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})] + \sum_{k=0}^{K_{\mathsf{max}}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}} , \tag{15}$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}\big[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] \leq \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}, \;\; \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}. \tag{16}$$

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

Proof We begin by recalling the definition

$$\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^k(\boldsymbol{\theta}). \tag{17}$$

Notice that

$$\begin{aligned} \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^{k+1})}, \{z_{i,m}^{(\tau_i^{k+1})}\}_{m=1}^{M_{(\tau_i^{k+1})}}) \\ &= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big). \end{aligned} \tag{18}$$

Furthermore, we recall that

$$\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \tfrac{1}{n} \textstyle\sum_{i=1}^{n} \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \widehat{e}^{(k)}(\boldsymbol{\theta}) := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \tag{19}$$

Due to S2, we have

$$\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \tag{20}$$

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

$$
\begin{aligned}
&\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) \leq \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\
&= \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \tfrac{1}{n}\big(\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})\big)
\end{aligned}
\tag{21}
$$

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

Need to improve upper bound here. H2 is too restricting

$$
\begin{aligned}
&\mathbb{E}_{i_k}\big[\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},\{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}})|\mathcal{F}_k,i_k\big]|\mathcal{F}_k\big] \\
&= \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathbb{E}_{i_k}\big[\mathbb{E}\big[\frac{1}{M_{(k)}}\sum_{m=1}^{M_{(k)}} r_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)},z_{i_k,m}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(k)})|\mathcal{F}_k,i_k\big]|\mathcal{F}_k\big] \\
&\leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}},
\end{aligned}
\tag{22}
$$

where the last inequality is due to H3. Moreover,

$$
\mathbb{E}\big[\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_{i_k}^k)},\{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})|\mathcal{F}_k\big] = \frac{1}{n}\sum_{i=1}^{n}\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)},\{z_{i,m}^{(\tau_i^k)}\}_{m=1}^{M_{(\tau_i^k)}}) = \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}).
\tag{23}
$$

Taking the conditional expectations on both sides of (21) and re-arranging terms give:

$$
\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}}
\tag{24}
$$

Proceeding from (24), we observe the following lower bound for the left hand side

$$
\begin{aligned}
&\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \overset{(a)}{=} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \\
&\overset{(b)}{\geq} \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2L}\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
&= \underbrace{\frac{1}{n}\sum_{i=1}^{n}\Big\{\frac{1}{M_{(\tau_i^k)}}\sum_{m=1}^{M_{(\tau_i^k)}} r_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)},z_{i,m}^{(\tau_i^k)}) - \widehat{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)};\boldsymbol{\theta}^{(\tau_i^k)})\Big\}}_{:=-\delta^{(k)}(\boldsymbol{\theta}^{(k)})} + \frac{1}{2L}\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2
\end{aligned}
\tag{25}
$$

where (a) is due to $\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$ [cf. S1], (b) is due to (20) and we have defined the summation in the last equality as $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$. Substituting the above into (24) yields

$$
\frac{\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \leq n\mathbb{E}\big[\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})|\mathcal{F}_k\big] + \frac{C_{\mathsf{r}}}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)})
\tag{26}
$$

Observe the following upper bound on the total expectations:

$$
\mathbb{E}\big[\delta^{(k)}(\boldsymbol{\theta}^{(k)})\big] \leq \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n}\frac{C_{\mathsf{r}}}{\sqrt{M_{(\tau_i^k)}}}\Big],
\tag{27}
$$

which is due to H3. 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 H3 and taking total expectations lead to

$$
\begin{aligned}
&\mathbb{E}\big[\|\nabla\widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] = \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}[\|\nabla\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{2LC_{\mathsf{r}}}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1}\mathbb{E}\Big[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n}\sum_{i=1}^{n}\frac{1}{\sqrt{M_{(\tau_i^k)}}}\Big]
\end{aligned}
\tag{28}
$$

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:

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^{k} \frac{1}{n}\left(1-\frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} \tag{29}$$

Taking the sum yields:

$$\begin{aligned}
&\sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{j=1}^{k} \frac{1}{n}\left(1-\frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\mathsf{max}}-1}\sum_{l=0}^{k-1} \frac{1}{n}\left(1-\frac{1}{n}\right)^{k-(l+1)} M_{(l)}^{-1/2} \\
&= \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2} \sum_{k=l+1}^{K_{\mathsf{max}}-1} \frac{1}{n}\left(1-\frac{1}{n}\right)^{k-(l+1)} \le \sum_{l=0}^{K_{\mathsf{max}}-1} M_{(l)}^{-1/2}
\end{aligned} \tag{30}$$

where the last inequality is due to upper bounding the geometric series. Plugging this back into (28) yields

$$\begin{aligned}
\mathbb{E}\big[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\big] &= \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}[\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\le \frac{2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})]}{K_{\mathsf{max}}} + \frac{1}{K_{\mathsf{max}}}\sum_{k=0}^{K_{\mathsf{max}}-1} \frac{4LC_{\mathsf{r}}}{\sqrt{M_{(k)}}} = \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}.
\end{aligned} \tag{31}$$

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

$$\begin{aligned}
g^{(k)} &= \inf_{\boldsymbol{\theta}\in\Theta} \frac{\mathcal{L}'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)})}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|} \\
&= \inf_{\boldsymbol{\theta}\in\Theta} \left\{ \frac{\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|} - \frac{\left\langle \nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) \,|\, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)} \right\rangle}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|} \right\} \\
&\ge -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta} \frac{\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)}-\boldsymbol{\theta}\|}
\end{aligned} \tag{32}$$

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

$$\begin{aligned}
&\frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\
&= \underbrace{\widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)})}_{\ge 0} - \widetilde{\mathcal{L}}^{(k)\prime}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}) + \frac{1}{n}\sum_{i=1}^{n} \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\
&\ge \frac{1}{n}\sum_{i=1}^{n} \left\{ \widehat{\mathcal{L}}_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) - \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r_i'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta}-\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) \right\}
\end{aligned} \tag{33}$$

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

$$g^{(k)} \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta}\in\Theta} (-\epsilon^{(k)}(\boldsymbol{\theta})) \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| - \sup_{\boldsymbol{\theta}\in\Theta} |\epsilon^{(k)}(\boldsymbol{\theta})|. \tag{34}$$

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

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

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

$$\mathbb{E}[g_-^{(K)}] \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|] + \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(K)}(\boldsymbol{\theta})] \tag{36}$$

We note that

$$\Big(\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|]\Big)^2 \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta(K_{\mathsf{max}})}{K_{\mathsf{max}}}, \tag{37}$$

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

$$\begin{aligned} \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(K)}(\boldsymbol{\theta})] &= \frac{1}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}} \mathbb{E}[\sup_{\boldsymbol{\theta}\in\Theta} \epsilon^{(k)}(\boldsymbol{\theta})] \overset{(a)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} \mathbb{E}\Big[\frac{1}{n}\sum_{i=1}^{n} M_{(\tau_i^k)}^{-1/2}\Big] \\ &\overset{(b)}{\leq} \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2} \end{aligned} \tag{38}$$

where (a) is due to H3 and (b) is due to (30). This implies

$$\mathbb{E}[g_-^{(K)}] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \sum_{k=0}^{K_{\mathsf{max}}-1} M_{(k)}^{-1/2}, \tag{39}$$

and concludes the proof of the theorem. □

4 Numerical examples

Have Logistic Regression with missing values and MNIST Bayesian LeNet5 from last year paper.

Working on CIFAR-10 using Bayesian ResNet.