

---

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

---

Anonymous Author(s)

Affiliation

Address

email

## Abstract

1 To be completed

## 2 1 Introduction

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

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

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

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

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

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

## 10 2 MISSO Algorithm

11 For any  $i \in \llbracket 1, n \rrbracket$ , we consider a surrogate function  $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}})$  which satisfies

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

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

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

14 **S2.** For any  $\bar{\boldsymbol{\theta}}_i \in \Theta$ ,  $i \in \llbracket 1, n \rrbracket$  and some  $\epsilon > 0$ , the difference function  $\widehat{e}(\boldsymbol{\theta}; \{\bar{\boldsymbol{\theta}}_i\}_{i=1}^n) :=$   
15  $\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \bar{\boldsymbol{\theta}}_i) - \mathcal{L}(\boldsymbol{\theta})$  is defined for all  $\boldsymbol{\theta} \in \Theta_\epsilon$  and differentiable for all  $\boldsymbol{\theta} \in \Theta$ , where  
16  $\Theta_\epsilon = \{\boldsymbol{\theta} \in \mathbb{R}^d, \inf_{\boldsymbol{\theta}' \in \Theta} \|\boldsymbol{\theta} - \boldsymbol{\theta}'\| < \epsilon\}$  is an  $\epsilon$ -neighborhood set of  $\Theta$ . Moreover, for some constant  
17  $L$ , the gradient satisfies

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

---

**Algorithm 1** MISSO method
 

---

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

$$\tilde{\mathcal{A}}_i^0(\theta) := \tilde{\mathcal{L}}_i(\theta; \theta^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(0)}}), \quad i \in \llbracket 1, n \rrbracket. \quad (7)$$

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

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

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

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

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

- 21 The MISSO method replaces the expectation in (5) by *Monte Carlo* integration and then optimizes
- 22 (1) incrementally.
- 23 Denote by  $M \in \mathbb{N}$  the Monte Carlo batch size and let  $z_m \in Z$ ,  $m = 1, \dots, M$  be a set of samples.
- 24 To this end, we define

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

- 25 and we summarize the proposed MISSO method in Algorithm 1.

### 3 Convergence Analysis

- 27 We provide non-asymptotic convergence bound for the MISSO method.

28 **H1.** For all  $i \in \llbracket 1, n \rrbracket$ ,  $\bar{\theta} \in \Theta$ ,  $z_i \in Z$ , the measurable function  $r_i(\theta; \bar{\theta}, z_i)$  is convex in  $\theta$  and is

29 lower bounded.

30 **H2.** For all  $i \in \llbracket 1, n \rrbracket$ ,  $(\theta, \bar{\theta}) \in \Theta^2$ ,  $z_i \in Z$  we assume the existence of an majorizing function

31  $m_r : Z \rightarrow \mathbb{R}$  and a constant  $C_r < \infty$  such that:

$$\sup_{M>0} \frac{1}{\sqrt{M}} \sum_{m=1}^M \left\{ r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta}) \right\} < m_r(z_i) \quad \text{and} \quad \mathbb{E}_{\bar{\theta}}[m_r(z_i)|\mathcal{F}] < C_r \quad (9)$$

- 32 where  $\mathcal{F}$  is the filtration of the total randomness and we denoted by  $\mathbb{E}_{\bar{\theta}}[\cdot]$  the expectation w.r.t. a
- 33 Markov chain  $\{z_{i,m}\}_{m=1}^M$  with initial distribution  $\xi_i(\cdot; \bar{\theta})$ , transition kernel  $P_{i,\bar{\theta}}$ , and stationary
- 34 distribution  $p_i(\cdot; \bar{\theta})$ . Besides,

$$\sup_{M>0} \frac{1}{\sqrt{M}} \sum_{m=1}^M \left\{ \frac{\hat{\mathcal{L}}'_i(\theta, \theta - \bar{\theta}; \bar{\theta}) - r'_i(\theta, \theta - \bar{\theta}; \bar{\theta}, z_{i,m})}{\|\bar{\theta} - \theta\|} \right\} < m_{\text{gr}}(z_i) \quad \text{and} \quad \mathbb{E}_{\bar{\theta}}[m_{\text{gr}}(z_i)|\mathcal{F}] < C_{\text{gr}} \quad (10)$$

**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  $v : Z \rightarrow \mathbb{R}$  and defined as  $v(z) := r_i(\theta; \bar{\theta}, z_{i,m}) - \hat{\mathcal{L}}_i(\theta; \bar{\theta})$  for all  $z \in Z$  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  $(X_i, i \in \llbracket 1, M \rrbracket)$  random variables taking values in  $(0, 1)$ , we have:

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

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

$$\mathbb{E} \sup_f |X_f| = \mathbb{E} \sup_f |X_f - X_0| \leq \frac{1}{\sqrt{M}} \int_0^1 \sqrt{\log \mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)} d\varepsilon \quad (12)$$

where  $\mathcal{N}(\mathcal{F}, \|\cdot\|_\infty, \varepsilon)$  is the bracketing number and  $\varepsilon$  denotes the level of approximation (the bracketing number goes to infinity when  $\varepsilon \rightarrow 0$ ). 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:

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

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:

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

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

Also, denote

$$\hat{\mathcal{L}}^{(k)}(\theta) := \frac{1}{n} \sum_{i=1}^n \hat{\mathcal{L}}_i(\theta; \theta^{(\tau_i^k)}), \quad \hat{e}^{(k)}(\theta) := \hat{\mathcal{L}}^{(k)}(\theta) - \mathcal{L}(\theta). \quad (15)$$

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

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

$$\Delta_{(K_{\max})} := 2nL\mathbb{E}[\tilde{\mathcal{L}}^{(0)}(\theta^{(0)}) - \tilde{\mathcal{L}}^{(K_{\max})}(\theta^{(K_{\max})})] + \sum_{k=0}^{K_{\max}-1} \frac{4LC_r}{\sqrt{M_{(k)}}}, \quad (16)$$

Then we have following non-asymptotic bounds:

$$\mathbb{E}[\|\nabla \hat{e}^{(K)}(\theta^{(K)})\|^2] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \quad (17)$$

$$\mathbb{E}[g_-(\theta^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}. \quad (18)$$

62 Note that  $\Delta_{(K_{\max})}$  is finite for any  $K_{\max} \in \mathbb{N}$ . As expected, the MISSO method converges to a  
 63 stationary point of (1) asymptotically and at a sublinear rate  $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\max}})$ .

64 Furthermore, we remark that the MISO method can be analyzed in Theorem 1 as a special case  
 65 of the MISSO method satisfying  $C_r = C_{gr} = 0$ . In this case, while the asymptotic convergence  
 66 is well known from [Mairal, 2015] [cf. H2], Eq. (17) gives a non-asymptotic rate of  $\mathbb{E}[g_-^{(K)}] \leq$   
 67  $\mathcal{O}(\sqrt{nL/K_{\max}})$  which is new to our best knowledge.

68 Next, we show that under an additional assumption on the sequence of batch size  $M_{(k)}$ , the MISSO  
 69 method converges almost surely to a stationary point:

70 **Theorem 2.** *Under S1, S2, H1, H2. In addition, assume that  $\{M_{(k)}\}_{k \geq 0}$  is a non-decreasing  
 71 sequence of integers which satisfies  $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$ . Then:*

- 72 1. *the negative part of the stationarity measure converges almost surely to zero,*  
 73 *i.e.,  $\lim_{k \rightarrow \infty} g_-(\theta^{(k)}) = 0$  a.s..*
- 74 2. *the objective value  $\mathcal{L}(\theta^{(k)})$  converges almost surely to a finite number  $\underline{\mathcal{L}}$ ,*  
 75 *i.e.,  $\lim_{k \rightarrow \infty} \mathcal{L}(\theta^{(k)}) = \underline{\mathcal{L}}$  a.s..*

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

## 78 4 Numerical Experiments

### 79 4.1 Binary logistic regression with missing values

80 This application follows **Example 1** described in Section 2. We consider a binary regression setup,  
 81  $((y_i, z_i), i \in \llbracket n \rrbracket)$  where  $y_i \in \{0, 1\}$  is a binary response and  $z_i = (z_{i,j} \in \mathbb{R}, j \in \llbracket p \rrbracket)$  is a covariate  
 82 vector. The vector of covariates  $z_i = [z_{i,\text{mis}}, z_{i,\text{obs}}]$  is not fully observed where we denote by  $z_{i,\text{mis}}$   
 83 the missing values and  $z_{i,\text{obs}}$  the observed covariate. It is assumed that  $(z_i, i \in \llbracket n \rrbracket)$  are i.i.d. and  
 84 marginally distributed according to  $\mathcal{N}(\beta, \Omega)$  where  $\beta \in \mathbb{R}^p$  and  $\Omega$  is a positive definite  $p \times p$  matrix.  
 85 We define the conditional distribution of the observations  $y_i$  given  $z_i = (z_{i,\text{mis}}, z_{i,\text{obs}})$  as:

$$p_i(y_i|z_i) = S(\delta^\top \bar{z}_i)^{y_i} (1 - S(\delta^\top \bar{z}_i))^{1-y_i} \quad (19)$$

86 where for  $u \in \mathbb{R}$ ,  $S(u) = 1/(1+e^{-u})$ ,  $\delta = (\delta_0, \dots, \delta_p)$  are the logistic parameters and  $\bar{z}_i = (1, z_i)$ .  
 87 We are interested in estimating  $\delta$  and finding the latent structure of the covariates  $z_i$ . Here,  $\theta =$   
 88  $(\delta, \beta, \Omega)$  is the parameter to estimate. For  $i \in \llbracket n \rrbracket$ , the complete data log-likelihood is expressed  
 89 as:

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

90 **MISSO update:** At the  $k$ -th iteration, and after the initialization, for all  $i \in \llbracket n \rrbracket$ , of the latent  
 91 variables  $(z_i^{(0)})$ , the MISSO algorithm consists in picking an index  $i_k$  uniformly on  $\llbracket n \rrbracket$ , complet-  
 92 ing the observations by sampling a Monte Carlo batch  $\{z_{i_k, \text{mis}, m}^{(k)}\}_{m=1}^{M(\tau_i^k)}$  of missing values from the  
 93 conditional distribution  $p(z_{i_k, \text{mis}}|z_{i_k, \text{obs}}, y_{i_k}; \theta^{(k-1)})$  using an MCMC sampler and computing the  
 94 estimated parameters as follows:

$$\begin{aligned} \beta^{(k)} &= \arg \min_{\beta \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega^{(k)}, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{M(\tau_i^k)} \sum_{m=1}^{M(\tau_i^k)} z_{i,m}^{(k)} \\ \Omega^{(k)} &= \arg \min_{\Omega \in \Theta} \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i^{(2)}(\beta^{(k)}, \Omega, \theta^{(\tau_i^k)}, \{z_{i,m}\}_{m=1}^{M(\tau_i^k)}) = \frac{1}{n} \sum_{i=1}^n \frac{1}{M(\tau_i^k)} \sum_{m=1}^{M(\tau_i^k)} z_{i,m}^{(k)} (z_{i,m}^{(k)})^\top - \beta^{(k)} (\beta^{(k)})^\top \\ \delta^{(k)} &= \frac{1}{n} \sum_{i=1}^n \delta^{(\tau_i^k)} - (\tilde{H}^{(k)})^{-1} \tilde{D}^{(k)}. \end{aligned} \quad (20)$$

95 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  
 96  $\tilde{D}^{(k)} = \frac{1}{n} \sum_{i=1}^n \tilde{D}_i^{(\tau_i^k)}$  and  $\tilde{H}^{(k)} = \frac{1}{n} \sum_{i=1}^n \tilde{H}_i^{(\tau_i^k)}$ . Besides,  $\tilde{\mathcal{L}}_i^{(1)}(\beta, \Omega, \bar{\theta}, \{z_m\}_{m=1}^M)$  and  
 97  $\tilde{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta}, \{z_m\}_{m=1}^M)$  are defined as MC approximation of  $\hat{\mathcal{L}}_i^{(1)}(\beta, \Omega, \bar{\theta})$  and  $\hat{\mathcal{L}}_i^{(2)}(\beta, \Omega, \bar{\theta})$ , for  
 98 all  $i \in \llbracket n \rrbracket$ .

99 See Appendix ?? for more explanation.

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

104 Similar to [Jiang et al., 2018], we select  $p = 16$  influential quantitative measurements, described  
 105 in Appendix ??, on  $n = 6384$  patients, and we adopt the logistic regression model with missing  
 106 covariates in (19) to predict the risk of a severe hemorrhage which is one of the main cause of  
 107 death after a major trauma. Note as the dataset considered is heterogeneous – coming from multiple  
 108 sources with frequently missed entries – we apply the latent data model described in the above.  
 109 For the Monte-Carlo sampling of  $z_{i,\text{mis}}$ , we run a Metropolis Hastings algorithm with the target  
 110 distribution  $p(\cdot|z_{i,\text{obs}}, y_i; \theta^{(k)})$  whose procedure is detailed in Appendix ??.

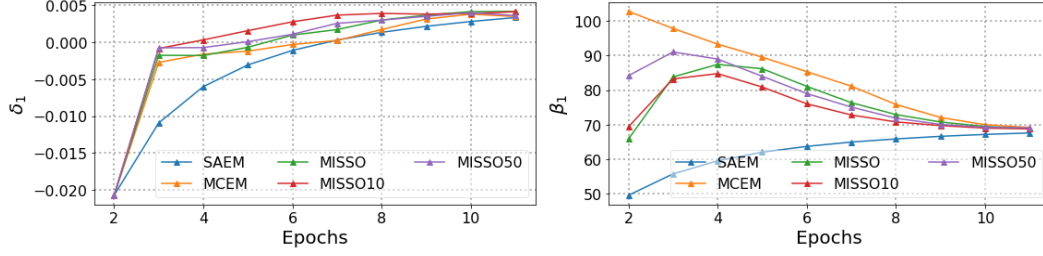


Figure 1: Convergence of first component of the vector of parameters  $\delta$  and  $\beta$  for the SAEM, the MCEM and the MISSO methods. The convergence is plotted against the number of passes over the data.

111 We compare in Figure 1 the convergence behavior of the estimated parameters  $\beta$  using SAEM  
 112 [Delyon et al., 1999] (with stepsize  $\gamma_k = 1/k$ ), MCEM [Wei and Tanner, 1990] and the proposed  
 113 MISSO method. For the MISSO method, we set the batch size to  $M_{(k)} = 10 + k^2$  and we examine  
 114 with selecting different number of functions in Line 5 in the method – the default settings with  
 115 1 function (MISSO), 10% (MISSO10) and 50% (MISSO50) of the functions per iteration. From  
 116 Figure 1, the MISSO method converges to a static value with less number of epochs than the MCEM,  
 117 SAEM methods. It is worth noting that the difference among the MISSO runs for different number  
 118 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  $\llbracket n \rrbracket$ ; **step (ii)** sample a Monte Carlo batch  $\{z_m^{(k)}\}_{m=1}^{M(k)}$  from  $\mathcal{N}(0, \mathbf{I})$ ; and **step (iii)** update the parameters as

$$\mu_\ell^{(k)} = \frac{1}{n} \sum_{i=1}^n \mu_\ell^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\mu_\ell, i}^{(k)} \quad \text{and} \quad \sigma^{(k)} = \frac{1}{n} \sum_{i=1}^n \sigma^{(\tau_i^k)} - \frac{\gamma}{n} \sum_{i=1}^n \hat{\delta}_{\sigma, i}^{(k)}, \quad (21)$$

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

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

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

**Bayesian LeNet-5 on MNIST [LeCun et al., 1998]:** 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,  $28 \times 28$  images. Each image is labelled with its corresponding number (from zero to nine). Under the prior distribution  $\pi$ , 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  $\mu = (\mu_\ell, \ell \in \llbracket d \rrbracket)$  where  $d$  is the number of weights in the neural network. We use the re-parametrization as  $w = t(\boldsymbol{\theta}, z) = \mu + \sigma z$  with  $z \sim \mathcal{N}(0, \mathbf{I})$ .

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

layer type	width	stride	padding	input shape	nonlinearity
convolution ( $5 \times 5$ )	6	1	0	$1 \times 32 \times 32$	ReLU
max-pooling ( $2 \times 2$ )		2	0	$6 \times 28 \times 28$	
convolution ( $5 \times 5$ )	6	1	0	$1 \times 14 \times 14$	ReLU
max-pooling ( $2 \times 2$ )		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 [He et al., 2016] on CIFAR-10 [Krizhevsky et al., 2012]:** We train here the Bayesian variant of the ResNet-18 neural network introduced in [He et al., 2016] on CIFAR-10. The latter dataset is composed of  $n = 60\,000$  handwritten digits,  $32 \times 32$  colour images in 10 classes, with 6 000 images per class. As in the previous example, the weights are assumed independent and identically distributed according to  $\mathcal{N}(0, 1)$ . The source code used as a backbone here can be found in the TensorFlow Probability Github repo ([https://github.com/tensorflow/probability/blob/master/tensorflow\\_probability/examples/cifar10\\_bnn.py](https://github.com/tensorflow/probability/blob/master/tensorflow_probability/examples/cifar10_bnn.py)) where the default hyperparameters, as the L annealing constant or the number of MC samples, were used for the benchmark methods. For better efficiency and lower variance, the Flipout estimator [?] is preferred than a simple reparametrization trick for ResNet-18.

layer type	Output Size	ResNet-18	nonlinearity
conv1	$112 \times 112 \times 64$	$7 \times 7, 64, \text{stride } 2$	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	$1 \times 1 \times 512$	$7 \times 7$ average pool	ReLU
fully connected	1000	$512 \times 1000$ fully connections	
softmax	1000		

Table 2: ResNet-18 architecture

**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  $M_{(k)} = k$ .

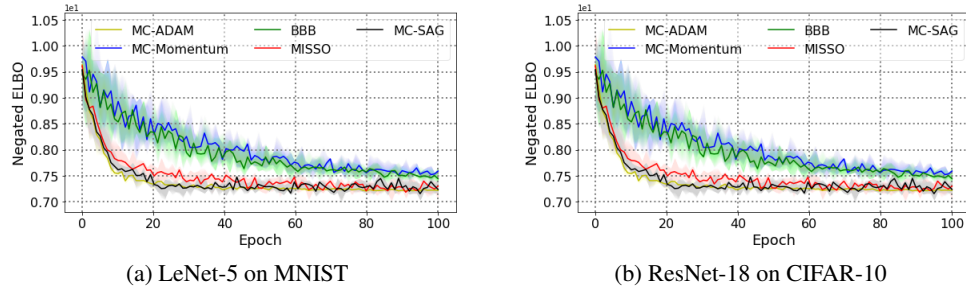


Figure 2: (a) Negated ELBO versus epochs elapsed for fitting the Bayesian LeNet-5 on MNIST using different algorithms. (b) ELBO versus epochs elapsed for fitting the Bayesian ResNet-18 on CIFAR-10 using different algorithms. The solid curve is obtained from averaging over 5 independent runs of the methods,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

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>.
- A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.
- 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.
- J. Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM J. Optim.*, 25(2):829–855, 2015. ISSN 1052-6234. doi: 10.1137/140957639. URL <https://doi.org/10.1137/140957639>.
- 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>.

## 206 A Proof of Theorem 1

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

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

209 Then we have following non-asymptotic bounds:

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

210 **Proof** We begin by recalling the definition

$$\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{A}}_i^k(\boldsymbol{\theta}). \quad (22)$$

211 Notice that

$$\begin{aligned} \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}) &= \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^{k+1})}, \{z_{i,m}^{(\tau_i^{k+1})}\}_{m=1}^{M_{(\tau_i^{k+1})}}) \\ &= \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) + \frac{1}{n} (\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})). \end{aligned} \quad (23)$$

212 Furthermore, we recall that

$$\hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) := \frac{1}{n} \sum_{i=1}^n \hat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(\tau_i^k)}), \quad \hat{e}^{(k)}(\boldsymbol{\theta}) := \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}) - \mathcal{L}(\boldsymbol{\theta}). \quad (24)$$

213 Due to S2, we have

$$\|\nabla \hat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \leq 2L\hat{e}^{(k)}(\boldsymbol{\theta}^{(k)}). \quad (25)$$

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

$$\begin{aligned} \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) &\leq \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k)}) \\ &= \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{n} (\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) - \tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})) \end{aligned} \quad (26)$$

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

216 We observe that the conditional expectation evaluates to

$$\begin{aligned} \mathbb{E}_{i_k} [\mathbb{E}[\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k,m}^{(k)}\}_{m=1}^{M_{(k)}}) | \mathcal{F}_k, i_k] | \mathcal{F}_k] \\ = \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \mathbb{E}_{i_k} [\mathbb{E}[\frac{1}{M_{(k)}} \sum_{m=1}^{M_{(k)}} r_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, z_{i_k,m}^{(k)}) - \hat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}) | \mathcal{F}_k, i_k] | \mathcal{F}_k] \\ \leq \mathcal{L}(\boldsymbol{\theta}^{(k)}) + \frac{C_r}{\sqrt{M_{(k)}}}, \end{aligned} \quad (27)$$

217 where the last inequality is due to H2. Moreover,

$$\mathbb{E}[\tilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k,m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}}) | \mathcal{F}_k] = \frac{1}{n} \sum_{i=1}^n \tilde{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, \{z_{i,m}^{(\tau_i^k)}\}_{m=1}^{M_{(\tau_i^k)}}) = \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}). \quad (28)$$

218 Taking the conditional expectations on both sides of (26) and re-arranging terms give:

$$\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \leq n\mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) | \mathcal{F}_k] + \frac{C_r}{\sqrt{M_{(k)}}} \quad (29)$$

219 Proceeding from (29), we observe the following lower bound for the left hand side

$$\begin{aligned}
& \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)}) \stackrel{(a)}{=} \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) \\
& \stackrel{(b)}{\geq} \tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) + \frac{1}{2L} \|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
& = \frac{1}{n} \sum_{i=1}^n \left\{ \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) - \hat{\mathcal{L}}_i(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \right\} + \frac{1}{2L} \|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2 \\
& \quad \underbrace{\hspace{10em}}_{:= -\delta^{(k)}(\boldsymbol{\theta}^{(k)})}
\end{aligned} \tag{30}$$

220 where (a) is due to  $\hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0$  [cf. S1], (b) is due to (25) and we have defined the summation in  
221 the last equality as  $-\delta^{(k)}(\boldsymbol{\theta}^{(k)})$ . Substituting the above into (29) yields

$$\frac{\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2}{2L} \leq n \mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) | \mathcal{F}_k] + \frac{C_r}{\sqrt{M_{(k)}}} + \delta^{(k)}(\boldsymbol{\theta}^{(k)}) \tag{31}$$

222 Observe the following upper bound on the total expectations:

$$\mathbb{E}[\delta^{(k)}(\boldsymbol{\theta}^{(k)})] \leq \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n \frac{C_r}{\sqrt{M_{(\tau_i^k)}}}\right], \tag{32}$$

223 which is due to H2. It yields

$$\mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \leq 2nL \mathbb{E}[\tilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \tilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})] + \frac{2LC_r}{\sqrt{M_{(k)}}} + \frac{1}{n} \sum_{i=1}^n \mathbb{E}\left[\frac{2LC_r}{\sqrt{M_{(\tau_i^k)}}}\right]$$

224 Finally, for any  $K_{\max} \in \mathbb{N}$ , we let  $K$  be a discrete r.v. that is uniformly drawn from  $\{0, 1, \dots, K_{\max} -$   
225  $1\}$ . Using H2 and taking total expectations lead to

$$\begin{aligned}
\mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] &= \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL \mathbb{E}[\tilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \tilde{\mathcal{L}}^{(K_{\max})}(\boldsymbol{\theta}^{(K_{\max})})]}{K_{\max}} + \frac{2LC_r}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\frac{1}{\sqrt{M_{(k)}}} + \frac{1}{n} \sum_{i=1}^n \frac{1}{\sqrt{M_{(\tau_i^k)}}}\right]
\end{aligned} \tag{33}$$

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

$$\mathbb{E}[M_{(\tau_i^k)}^{-1/2}] = \sum_{j=1}^k \frac{1}{n} \left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} \tag{34}$$

228 Taking the sum yields:

$$\begin{aligned}
\sum_{k=0}^{K_{\max}-1} \mathbb{E}[M_{(\tau_i^k)}^{-1/2}] &= \sum_{k=0}^{K_{\max}-1} \sum_{j=1}^k \frac{1}{n} \left(1 - \frac{1}{n}\right)^{j-1} M_{(k-j)}^{-1/2} = \sum_{k=0}^{K_{\max}-1} \sum_{l=0}^{k-1} \frac{1}{n} \left(1 - \frac{1}{n}\right)^{k-(l+1)} M_{(l)}^{-1/2} \\
&= \sum_{l=0}^{K_{\max}-1} M_{(l)}^{-1/2} \sum_{k=l+1}^{K_{\max}-1} \frac{1}{n} \left(1 - \frac{1}{n}\right)^{k-(l+1)} \leq \sum_{l=0}^{K_{\max}-1} M_{(l)}^{-1/2}
\end{aligned} \tag{35}$$

229 where the last inequality is due to upper bounding the geometric series. Plugging this back into (33)  
230 yields

$$\begin{aligned}
\mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] &= \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}[\|\nabla \hat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})\|^2] \\
&\leq \frac{2nL \mathbb{E}[\tilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \tilde{\mathcal{L}}^{(K_{\max})}(\boldsymbol{\theta}^{(K_{\max})})]}{K_{\max}} + \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \frac{4LC_r}{\sqrt{M_{(k)}}} = \frac{\Delta_{(K_{\max})}}{K_{\max}}.
\end{aligned} \tag{36}$$

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

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

$$\begin{aligned} g^{(k)} &= \inf_{\boldsymbol{\theta} \in \Theta} \frac{\mathcal{L}'(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \\ &= \inf_{\boldsymbol{\theta} \in \Theta} \left\{ \frac{\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} - \frac{\langle \nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) | \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)} \rangle}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \right\} \\ &\geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta} \in \Theta} \frac{\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)})}{\|\boldsymbol{\theta}^{(k)} - \boldsymbol{\theta}\|} \end{aligned} \quad (37)$$

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

$$\begin{aligned} &\frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\ &= \underbrace{\widetilde{\mathcal{L}}^{(k)'}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}) - \widetilde{\mathcal{L}}^{(k)'}(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)})}_{\geq 0} + \frac{1}{n} \sum_{i=1}^n \widehat{\mathcal{L}}'_i(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) \\ &\geq \frac{1}{n} \sum_{i=1}^n \left\{ \widehat{\mathcal{L}}'_i(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}) - \frac{1}{M_{(\tau_i^k)}} \sum_{m=1}^{M_{(\tau_i^k)}} r'_i(\boldsymbol{\theta}^{(k)}, \boldsymbol{\theta} - \boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_i^k)}, z_{i,m}^{(\tau_i^k)}) \right\} \end{aligned} \quad (38)$$

237 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  
 238 a scaled version of the above term as:

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

239 We have

$$g^{(k)} \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \inf_{\boldsymbol{\theta} \in \Theta} (-\epsilon^{(k)}(\boldsymbol{\theta})) \geq -\|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| - \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})|. \quad (39)$$

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

$$g_-^{(k)} \leq \|\nabla \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| + \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})|. \quad (40)$$

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

$$\mathbb{E}[g_-^{(K)}] \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|] + \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \epsilon^{(K)}(\boldsymbol{\theta})] \quad (41)$$

243 We note that

$$\left( \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|] \right)^2 \leq \mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta(K_{\max})}{K_{\max}}, \quad (42)$$

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

$$\begin{aligned} \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \epsilon^{(K)}(\boldsymbol{\theta})] &= \frac{1}{K_{\max}} \sum_{k=0}^{K_{\max}} \mathbb{E}[\sup_{\boldsymbol{\theta} \in \Theta} \epsilon^{(k)}(\boldsymbol{\theta})] \stackrel{(a)}{\leq} \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n M_{(\tau_i^k)}^{-1/2}\right] \\ &\stackrel{(b)}{\leq} \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2} \end{aligned} \quad (43)$$

245 where (a) is due to H2 and (b) is due to (35). This implies

$$\mathbb{E}[g_-^{(K)}] \leq \sqrt{\frac{\Delta(K_{\max})}{K_{\max}}} + \frac{C_{\text{gr}}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}, \quad (44)$$

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

## B Proof of Theorem 2

**Theorem.** Under S1, S2, H1, H2. In addition, assume that  $\{M_{(k)}\}_{k \geq 0}$  is a non-decreasing sequence of integers which satisfies  $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$ . Then:

1. the negative part of the stationarity measure converges almost surely to zero, i.e.,  $\lim_{k \rightarrow \infty} g_{-}(\boldsymbol{\theta}^{(k)}) = 0$  a.s..
2. the objective value  $\mathcal{L}(\boldsymbol{\theta}^{(k)})$  converges almost surely to a finite number  $\underline{\mathcal{L}}$ , i.e.,  $\lim_{k \rightarrow \infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$  a.s..

**Proof** We apply the following auxiliary lemma which proof can be found in Appendix C 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  $k \geq 1$ :

$$V_k \leq V_{k-1} - X_{k-1} + E_{k-1} \quad (45)$$

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  $V_{\infty}$ .

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

(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 (26) by re-arranging terms and observing that

$$\begin{aligned} \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) &\leq \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad - (\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k, m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})) \end{aligned} \quad (46)$$

Our idea is to apply Lemma 1. Under S1, 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

$$V_k := \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \inf_{k \geq 0} c_k \geq 0 \quad (47)$$

is a non-negative random variable.

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

$$X_k := \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(\tau_{i_k}^k)}; \boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \geq 0. \quad (48)$$

Thirdly, we define

$$\begin{aligned} E_k &= -(\widetilde{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)}) - \widehat{\mathcal{L}}^{(k+1)}(\boldsymbol{\theta}^{(k+1)})) + (\widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k, m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})) \\ &\quad + \frac{1}{n} (\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}})). \end{aligned} \quad (49)$$

Note that from the definitions (47), (48), (49), we have  $V_{k+1} \leq V_k - X_k + E_k$  for any  $k \geq 1$ .

Under H2, we observe that

$$\mathbb{E}[|\widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)}, \{z_{i_k, m}^{(k)}\}_{m=1}^{M_{(k)}}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})|] \leq C_r M_{(k)}^{-1/2} \quad (50)$$

$$\mathbb{E} \left[ \left| \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widetilde{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}, \{z_{i_k, m}^{(\tau_{i_k}^k)}\}_{m=1}^{M_{(\tau_{i_k}^k)}}) \right| \right] \leq C_r \mathbb{E} \left[ M_{(\tau_{i_k}^k)}^{-1/2} \right] \quad (51)$$

$$\mathbb{E} \left[ \left| \widetilde{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) \right| \right] \leq \frac{1}{n} \sum_{i=1}^n C_r \mathbb{E} \left[ M_{(\tau_i^k)}^{-1/2} \right] \quad (52)$$

Therefore,

$$\mathbb{E} [|E_k|] \leq \frac{C_r}{n} \left( M_{(k)}^{-1/2} + \mathbb{E} \left[ M_{(\tau_{i_k}^k)}^{-1/2} + \sum_{i=1}^n \{ M_{(\tau_i^k)}^{-1/2} + M_{(\tau_{i+1}^k)}^{-1/2} \} \right] \right) \quad (53)$$

Using (35) and the assumption on the sequence  $\{M_{(k)}\}_{k \geq 0}$ , we obtain that

$$\sum_{k=0}^{\infty} \mathbb{E} [|E_k|] < \frac{C_r}{n} (2 + 2n) \sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty. \quad (54)$$

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

$$\begin{aligned} \infty &> \sum_{k=0}^{\infty} \mathbb{E} [X_k] = \frac{1}{n} \sum_{k=0}^{\infty} \mathbb{E} [\widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(\tau_{i_k}^k)}) - \widehat{\mathcal{L}}_{i_k}(\boldsymbol{\theta}^{(k)}; \boldsymbol{\theta}^{(k)})] \\ &= \frac{1}{n} \sum_{k=0}^{\infty} \mathbb{E} [\widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) - \mathcal{L}(\boldsymbol{\theta}^{(k)})] = \frac{1}{n} \sum_{k=0}^{\infty} \mathbb{E} [\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})] \end{aligned} \quad (55)$$

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

$$\lim_{k \rightarrow \infty} \widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)}) = 0 \quad \text{a.s.} \quad (56)$$

and subsequently applying (25), we have  $\lim_{k \rightarrow \infty} \|\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})\| = 0$  almost surely. Finally, it follows from (25) and (40) that

$$\lim_{k \rightarrow \infty} g_-^{(k)} \leq \lim_{k \rightarrow \infty} \sqrt{2L} \sqrt{\widehat{e}^{(k)}(\boldsymbol{\theta}^{(k)})} + \lim_{k \rightarrow \infty} \sup_{\boldsymbol{\theta} \in \Theta} |\epsilon^{(k)}(\boldsymbol{\theta})| = 0, \quad (57)$$

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  $\lim_{k \rightarrow \infty} \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \underline{\mathcal{L}}$ . Applying (56) implies that

$$\underline{\mathcal{L}} = \lim_{k \rightarrow \infty} \widehat{\mathcal{L}}^{(k)}(\boldsymbol{\theta}^{(k)}) = \lim_{k \rightarrow \infty} \mathcal{L}(\boldsymbol{\theta}^{(k)}) \quad \text{a.s.} \quad (58)$$

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

## C 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  $V_{\infty}$ .

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

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

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

$$0 \leq V_k \leq V_0 - \sum_{j=1}^k X_j + \sum_{j=1}^k E_j \leq V_0 + \sum_{j=1}^k E_j \quad (59)$$

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

296 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  
 297  $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$ , the series  $\sum_{j=1}^{\infty} E_j$  converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \quad (60)$$

298 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:

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \end{aligned} \quad (61)$$

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

$$\mathbb{E}[\liminf_{k \rightarrow \infty} |W_k|] = \mathbb{E}[|W_{\infty}|] \leq \liminf_{k \rightarrow \infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \quad (62)$$

303 showing that the random variable  $W_{\infty}$  is integrable.

304 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  
 305  $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$  and by the monotone convergence theorem, we get:

$$\lim_{k \rightarrow \infty} \mathbb{E}[U_k] = \mathbb{E}\left[\lim_{k \rightarrow \infty} U_k\right] \quad (63)$$

306 Finally, we have:

$$\lim_{k \rightarrow \infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_{\infty} \quad \text{and} \quad \mathbb{E}\left[\lim_{k \rightarrow \infty} U_k\right] = \mathbb{E}[W_0] - \mathbb{E}[W_{\infty}] \quad (64)$$

307 showing that  $\mathbb{E}[W_{\infty}] = w_{\infty}$  and concluding the proof of (ii). Moreover, using (61) we have that  
 308  $W_k \leq W_{k-1} - X_k$  which yields:

$$\begin{aligned} \sum_{j=1}^{\infty} X_j &\leq W_0 - W_{\infty} < \infty \\ \sum_{j=1}^{\infty} \mathbb{E}[X_j] &\leq \mathbb{E}[W_0] - w_{\infty} < \infty \end{aligned} \quad (65)$$

309 which concludes the proof of the lemma.  $\square$