Minimization by Incremental Stochastic Surrogate Optimization for Large Scale Nonconvex Problems

Belhal Karimi[1], Hoi-To Wai[2], Eric Moulines[3] and Ping Li[1]

Baidu Research[1], Chinese University of Hong Kong[2], Ecole Polytechnique[3]

belhalkarimi@baidu.com, htwai@se.cuhk.edu.hk, eric.moulines@polytechnique.edu, liping11@baidu.com@gmail.com

Large Scale Optimization

- **Objective:** *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 $\mathcal{L}_i : \mathbb{R}^p \to \mathbb{R}$ is bounded from below and is (possibly) nonconvex and include a nonsmooth penalty.

- The gap $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)$ is L-smooth. Denote by $\langle \cdot | \cdot \rangle$ the scalar product, the stationary point condition is:

Definition 1. (Asymptotic Stationary Point Condition)
A sequence $(\theta^k)_{k\geq 0}$ satisfies the asymptotic stationary point condition if

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

Majorization-Minimization Scheme

- The MISO method (Mairal, 2015)

Algorithm 2 The MISO method (Mairal, 2015).

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

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

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

An Inctractablity for Latent Data Models

- Case when the surrogate functions computed in Algorithm **?? are not tractable**.
- Assume that the surrogate can be expressed as an integral over a set of latent variables $z = (z_i \in Z, i \in [n]) \in Z[]$.

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \int_Z r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i) p_i(z_i; \overline{\boldsymbol{\theta}}) \mu_i(dz_i) \quad \forall\, (\boldsymbol{\theta}, \overline{\boldsymbol{\theta}}) \in \Theta \times \Theta \,. \tag{3}$$

- Our scheme is based on the computation, at each iteration, of stochastic auxiliary functions for a mini-batch of components. For $i \in [n]$, the auxiliary function, noted $\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M)$ is a Monte Carlo approximation of the surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ defined by (3Doc-Start) such that:

$$\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M) := \frac{1}{M}\sum_{m=1}^{M} r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_m) \,, \tag{4}$$

where $\{z_i^m\}_{m=0}^{M-1}$ is a Monte Carlo batch.

MISSO Method

Algorithm 2 The MISSO method.

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

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

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

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

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

Global Convergence Analysis

Assumptions: we need a few regularity conditions in this case,

H 1. For all $i \in [n]$ and $\overline{\boldsymbol{\theta}} \in \Theta$, $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ is convex *w.r.t.* $\boldsymbol{\theta}$, and it holds $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta})$, $\forall\, \boldsymbol{\theta} \in \Theta$ where the equality holds when $\boldsymbol{\theta} = \overline{\boldsymbol{\theta}}$.

H2. For any $\overline{\boldsymbol{\theta}}_i \in \Theta$, $i \in [n]$ 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 $\|\nabla \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)\|^2 \leq 2L\, \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)$, $\forall\, \boldsymbol{\theta} \in \Theta$.

H3. For all $i \in [n]$, $\overline{\boldsymbol{\theta}} \in \Theta$, $z_i \in Z$, $r_i(\cdot; \overline{\boldsymbol{\theta}}, z_i)$ is convex on Θ and is lower bounded.

H4. For the samples $\{z_{i,m}\}_{m=1}^M$, there exist finite constants C_r and C_{gr} such that for all $i \in [n]$,

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

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

where 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 $\Pi_{i,\overline{\boldsymbol{\theta}}}$, and stationary distribution $p_i(\cdot; \overline{\boldsymbol{\theta}})$.

Theorem 1 *Under H1-H4. For any $K_{\max} \in \mathbb{N}$, let K be an independent discrete r.v. drawn uniformly from $\{0, ..., K_{\max}-1\}$ and define the following quantity:*

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

Then we have following non-asymptotic bounds:

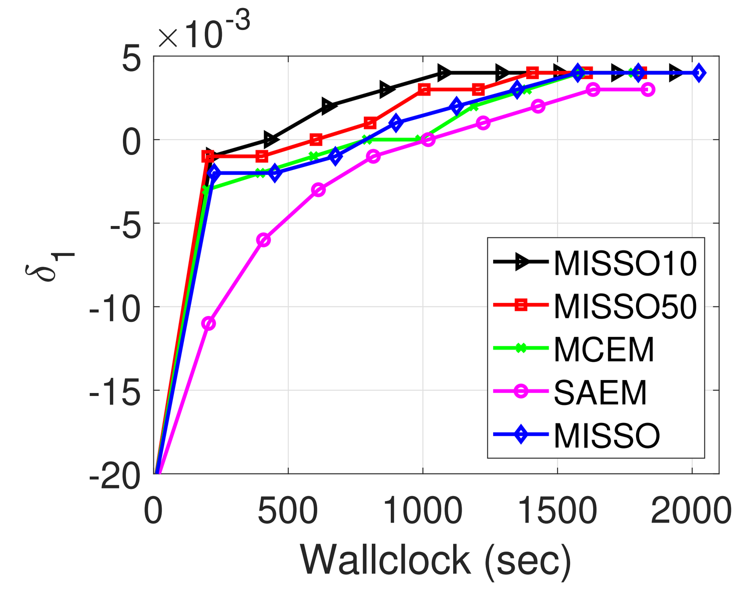
$$\mathbb{E}[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2] \leq \frac{\Delta_{(K_{\max})}}{K_{\max}} \text{ and } \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{gr}}{K_{\max}} \overline{M}_{(K_{\max})} \,. \tag{16}$$

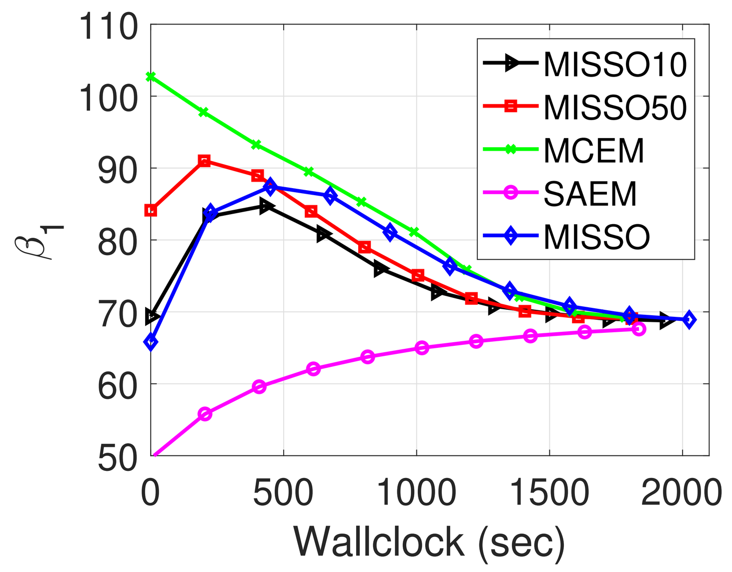
Numerical Experiments

- Logistic Regression **with missing values** on Traumabase (severe hemorrhage):

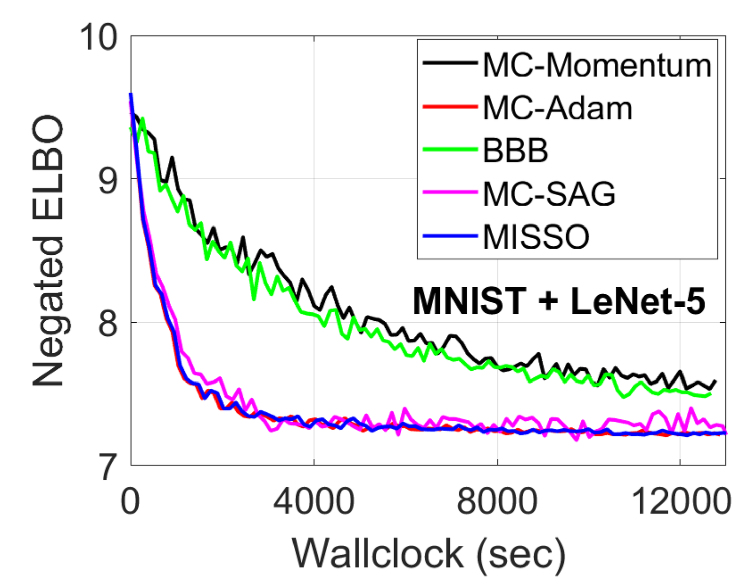
$$p_i(y_i|z_i) = S(\boldsymbol{\delta}^\top \bar{z}_i)^{y_i} \left(1 - S(\boldsymbol{\delta}^\top \bar{z}_i)\right)^{1-y_i} \,,$$

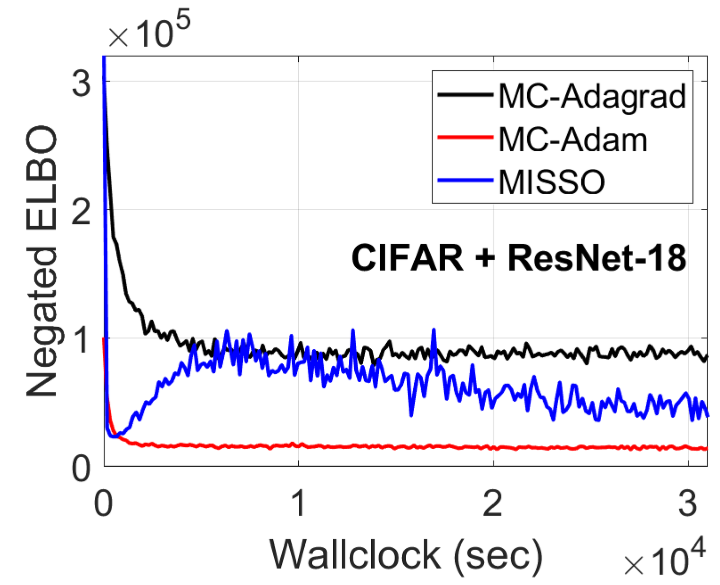
- 16 quantitative measurements, like BMI, age, blood pressure, heart rate at different stages after the accident on 6384 patients
- **MISSO is an incremental MCEM in this case**





- Bayesian variants of LeNet-5 and ResNet-18 on MNIST and CIFAR10:
- Variational inference and the ELBO loss to fit Bayesian Neural Networks on different datasets.
- **MISSO is an incremental VI in this case**





Conclusion

- Theorem 1 & 2 show the non-asymptotic convergence rate of biased SA scheme with smooth (possibly non-convex) Lyapunov function.
- With appropriate step size, in n iterations the SA scheme finds $\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] = \mathcal{O}(c_0 + \log n/\sqrt{n})$, where c_0 is the bias and $h(\cdot)$ is the mean field.
- Applications to online EM and online policy gradient.

References

Julien Mairal. Incremental majorization-minimization optimization with application to large-scale machine learning. *SIAM Journal on Optimization*, 25(2):829–855, 2015.