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:

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

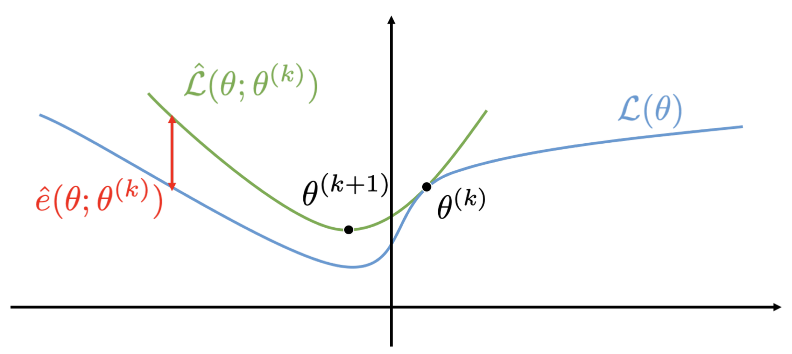
where $\mathcal{L}_i : \mathbb{R}^p \to \mathbb{R}$ is bounded from below and is (possibly) nonconvex and include a nonsmooth penalty.

- The gap $\widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)$ plays a key role in the convergence analysis and we require this error to be L-smooth for some constant $L > 0$ Denote by $\langle \cdot | \cdot \rangle$ the scalar product, we also introduce the following stationary point condition:

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



Algorithm 2 The MISO method (Mairal, 2015).

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

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

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

- **MISO Method**: fix any $n \geq 1$, we stop the SA at a random iteration N with

An Inctractability 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 (3) 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

- dd

Algorithm 2 The 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

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

for $k = 0, 1, ..., K_{\max}$ **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:

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

 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

Global Convergence Analysis

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

1. There exists a Borel measurable function $\hat{H} : \mathcal{H} \times \mathsf{X} \to \mathcal{H}$,

$$\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) = H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta}),\ \forall\, \boldsymbol{\eta} \in \mathcal{H}, x \in \mathsf{X}.$$

$\Longrightarrow$ existence of solution to the *Poisson equation*.

2. For all $\boldsymbol{\eta} \in \mathcal{H}$ and $x \in \mathsf{X}$, $\|\hat{H}_{\boldsymbol{\eta}}(x)\| \leq L_{PH}^{(0)}$, $\|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x)\| \leq L_{PH}^{(0)}$, and

$$\sup_{x \in \mathsf{X}} \|P_{\boldsymbol{\eta}}\hat{H}_{\boldsymbol{\eta}}(x) - P_{\boldsymbol{\eta}'}\hat{H}_{\boldsymbol{\eta}'}(x)\| \leq L_{PH}^{(1)}\|\boldsymbol{\eta} - \boldsymbol{\eta}'\|,\ \forall\, (\boldsymbol{\eta}, \boldsymbol{\eta}') \in \mathcal{H}^2.$$

$\Longrightarrow$ *smoothness* of $\hat{H}_{\boldsymbol{\eta}}(x)$, satisfied if $P_{\boldsymbol{\eta}}$, $H_{\boldsymbol{\eta}}(X)$ are smooth *w.r.t.* $\boldsymbol{\eta}$.

3. It holds that $\sup_{\boldsymbol{\eta} \in \mathcal{H}, x \in \mathsf{X}} \|H_{\boldsymbol{\eta}}(x) - h(\boldsymbol{\eta})\| \leq \sigma$.

$\Longrightarrow$ requires the noise is *uniformly bounded* for all $x \in \mathsf{X}$.

Example: assumptions 1 & 2 are satisfied if the Markov kernel P_{η_n} is geometrically ergodic + smooth, and the drift term is smooth *w.r.t.* $\boldsymbol{\eta}$.

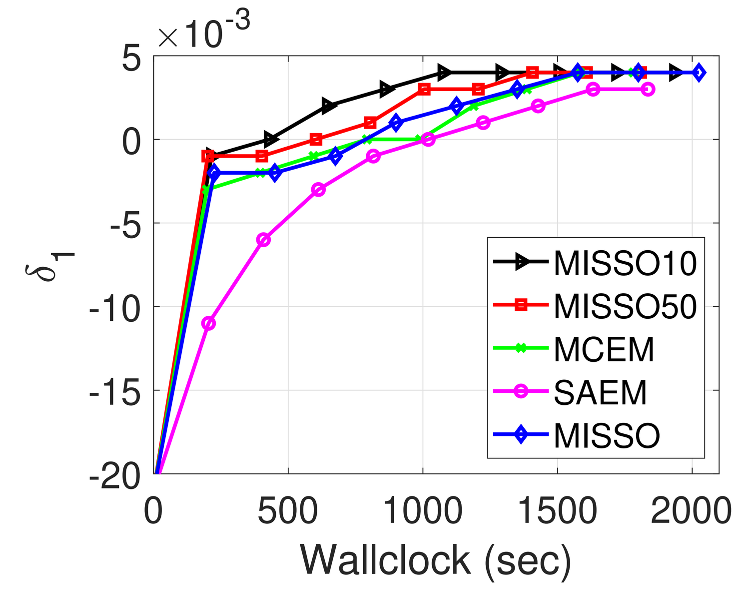
Theorem 1. Suppose that the step sizes are decreasing and $\gamma_1 \leq 0.5\big(c_1(L + C_h)\big)^{-1}$ (+other conditions). Let $V_{0,n} := \mathbb{E}[V(\boldsymbol{\eta}_0) - V(\boldsymbol{\eta}_{n+1})]$,

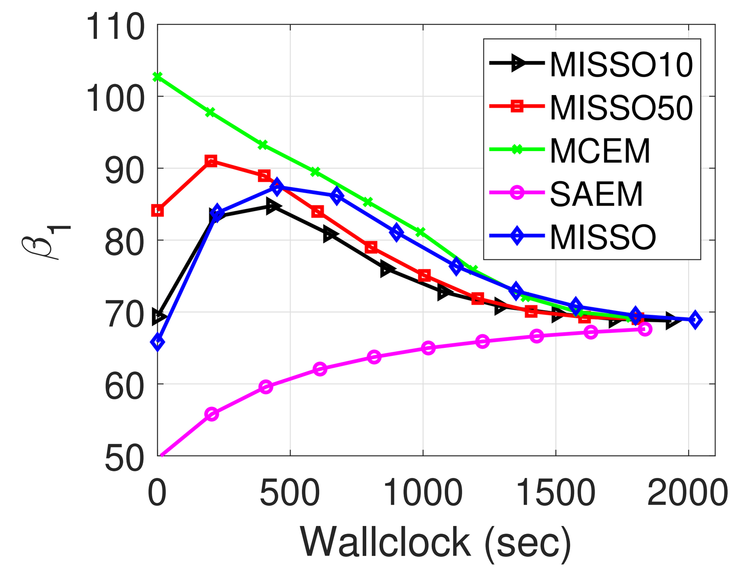
$$\mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] \leq \frac{2c_1\big(V_{0,n} + C_{0,n} + (\sigma^2 L + C_\gamma)\sum_{k=0}^n \gamma_{k+1}^2\big)}{\sum_{k=0}^n \gamma_{k+1}} + 2c_0\,.$$

- Set $\gamma_k = (2c_1 L(1 + C_h)\sqrt{k})^{-1} \Longrightarrow \mathbb{E}[\|h(\boldsymbol{\eta}_N)\|^2] = \mathcal{O}(c_0 + \log n/\sqrt{n})$ (same as Case 1).
- **Proof idea:** challenge is that e_{n+1} is not zero-mean $\Longrightarrow$ bound the sum of $\mathbb{E}[\langle \nabla V(\boldsymbol{\eta}_n) \,|\, e_{n+1} \rangle]$ w/ Poisson equation + a novel decomposition (cf. *Lemma 2*).

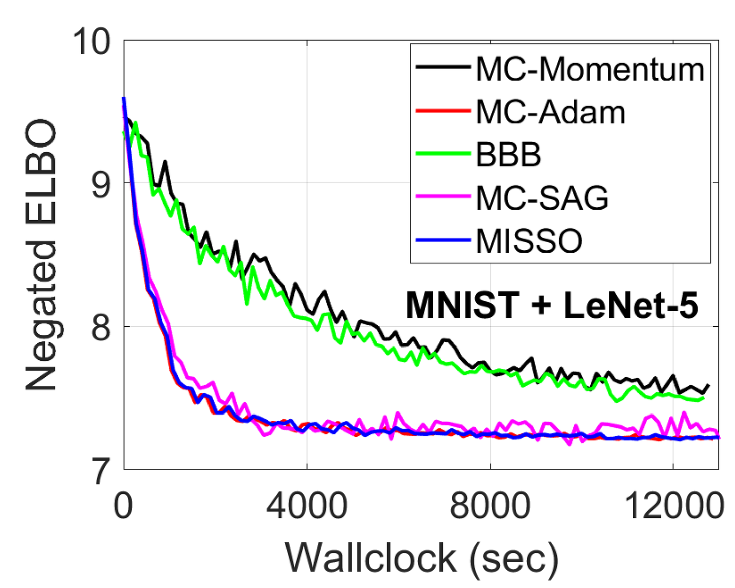
Numerical Experiments

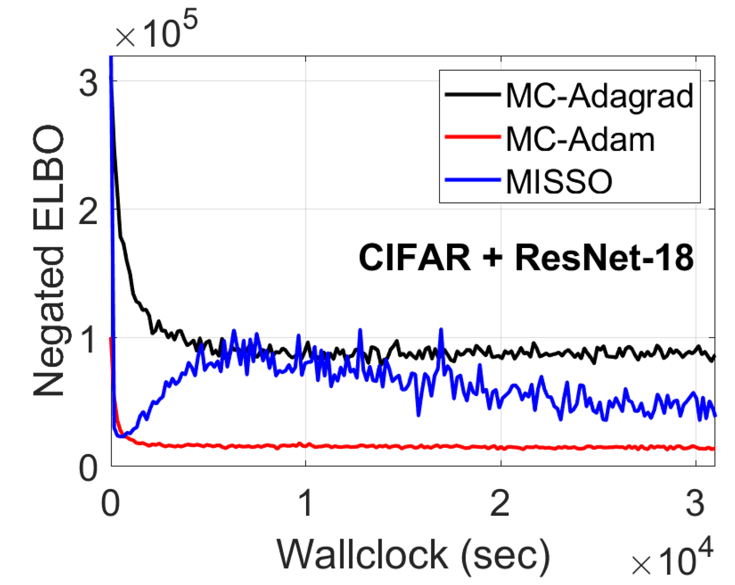
- Logistic Regression on Traumabase dataset (severe hemorrhage):





- Bayesian variants of LeNet-5 and ResNet-18 on MNIST and CIFAR10:





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