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]

Large Scale Optimization

• **Objective:** *Constrained* minimization problem of a finite sum of functions:

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

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.

Asymptotic Stationary Point Condition:

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

The MISO Method (Mairal, 2015)

• MISO has been proposed in (Mairal, 2015)

• Initialize the majorizing surrogate functions

$$\mathcal{A}_i^0(\boldsymbol{\theta}) := \widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \boldsymbol{\theta}^{(0)}), \; i \in [n].$$

• For $k > 0$:

• Pick i_k uniformly from $[n]$.

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

• **Return**: $\boldsymbol{\theta}^{(k+1)}$.

An Intractability for Latent Data Models

• Case when the surrogate functions computed in MISO **are not tractable**. • Surrogate is expressed as an integral over a set of latent variables z.

$$\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) := \int_{\mathsf{Z}} r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_i) p_i(z_i; \overline{\boldsymbol{\theta}}) \mu_i(dz_i) \; . \qquad (2)$$

• Our scheme is based on the computation of $\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M)$, a Monte Carlo approximation of the surrogate function $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}})$ defined by (2) such that:

$$\widetilde{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, \{z_m\}_{m=1}^M) := \frac{1}{M}\sum_{m=1}^{M} r_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}, z_m) \; , \qquad (3)$$

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

MISSO: Minimization by Incremental Stochastic Surrogate Optimization

Idea: We replace the expectation in (2) by a *Monte Carlo* integration and then optimizes the objective function in an incremental manner..

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

How does this Monte Carlo Approximation affects the convergence rate?

Global Convergence Analysis

• We bound the **Monte Carlo noise** by some constants using the notion of metric entropy (or bracketing number) see Van der Vaart (2000): $\mathbb{E}\left[\sup_{f\in\mathcal{F}} \left|\frac{1}{M}\sum_{i=1}^M f(z_{i,m}) - \mathbb{E}[f(z_i)]\right|\right] \leq \frac{CL}{\sqrt{M}}$.

• **Convergence Analysis**: Assume $\widehat{\mathcal{L}}_i(\boldsymbol{\theta}; \overline{\boldsymbol{\theta}}) \geq \mathcal{L}_i(\boldsymbol{\theta})$ and the error gap follows $\|\nabla \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)\|^2 \leq 2L\, \widehat{e}(\boldsymbol{\theta}; \{\overline{\boldsymbol{\theta}}_i\}_{i=1}^n)$

Theorem (MISSO Finite-Time Convergence) For any $K_{\mathsf{max}} \geq 1$, $K \sim \mathcal{U}([0, K_{\mathsf{max}} - 1])$ independent of the $\{i_k\}_{k=0}^{K_{\mathsf{max}}}$, we have the **global rate**:

$$\mathbb{E}\left[\|\nabla \widehat{e}^{(K)}(\boldsymbol{\theta}^{(K)})\|^2\right] \leq \frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}} \quad \text{and} \quad \mathbb{E}[g_-(\boldsymbol{\theta}^{(K)})] \leq \sqrt{\frac{\Delta_{(K_{\mathsf{max}})}}{K_{\mathsf{max}}}} + \frac{C_{\mathsf{gr}}}{K_{\mathsf{max}}} \overline{M}_{(K_{\mathsf{max}})} \; . \qquad (4)$$

where

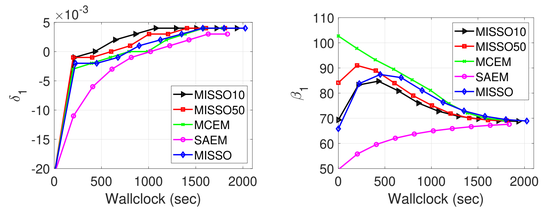
$$\Delta_{(K_{\mathsf{max}})} := 2nL\mathbb{E}[\widetilde{\mathcal{L}}^{(0)}(\boldsymbol{\theta}^{(0)}) - \widetilde{\mathcal{L}}^{(K_{\mathsf{max}})}(\boldsymbol{\theta}^{(K_{\mathsf{max}})})] + 4LC_{\mathsf{r}}\overline{M}_{(K_{\mathsf{max}})} \; .$$

Logistic Regression on Hemorrhage dataset

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

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

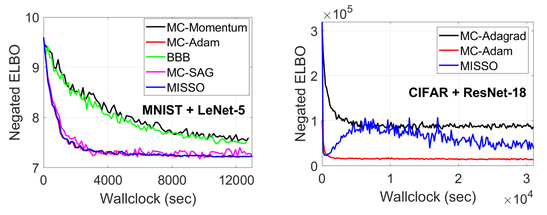
• 16 quantitative measurements, like BMI, age, blood pressure, heart rate at different stages after the accident on 6384 patients • **MISSO is an incremental MCEM in this case**



Bayesian Neural Networks using MISSO

• **Goal**: Train 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**



References

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

Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

The 33rd International Conference on Algorithmic Learning Theory, Paris, France