The 33rd International Conference on Algorithmic Learning Theory

Minimization by Incremental Stochastic Surrogate Optimization for Large Scale Nonconvex Problems

Belhal Karimi, Hoi-To Wai, Eric Moulines and Ping Li

Outline of the Talk

1. Large Scale ML and MM Scheme
2. MISSO: Algorithm and Global Convergence
3. Numerical Runs

Large-Scale ML

Constrained Optimization of Finite Sum

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

- Large finite-sum objective function
- Convex and compact Θ subset of $\mathbb{R}^d$
- Function $\mathcal{L}_i : \mathbb{R}^d \to \mathbb{R}$ is bounded from below and possibly non convex and nonsmooth

Some examples

- Maximum likelihood estimation

$$\mathcal{L}_i(\theta) := -\log p_i(y_i, \theta)$$

- Variational Inference

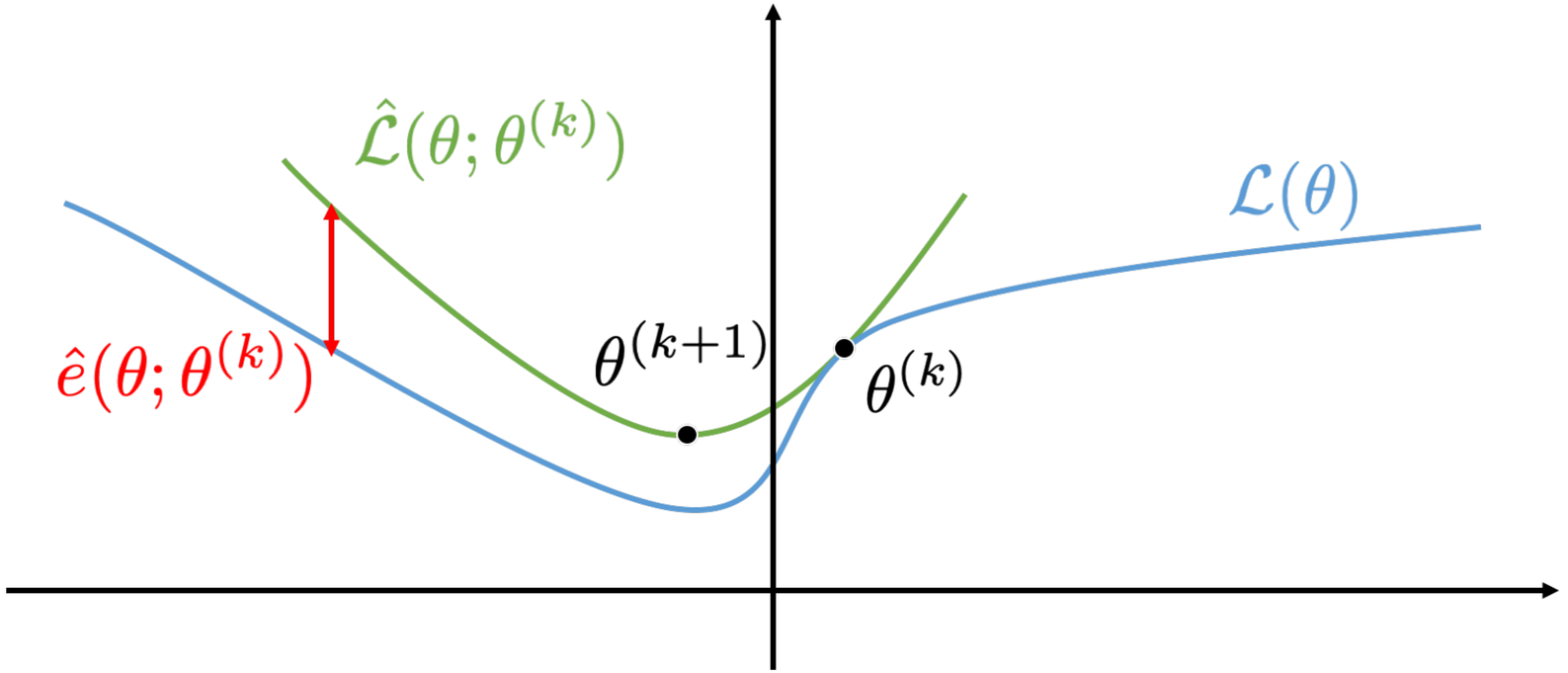
$$\mathcal{L}_i(\theta) := \mathrm{KL}\left(q(w;\theta) \| p_i\left(w|y_i, x_i\right)\right)$$

- Logistic Regression {-1/1} binary outputs

$$\mathcal{L}_i(\theta) := \log(1 + e^{-y_i \theta^\top x_i})$$

Majorization-Minimization Principle

[Lange, 2013]



- Iteratively minimize locally tight upper bounds on the objective
- Drives the objective function downwards
- The approximation error $\hat{e}(\theta,\bar{\theta})$ at $\bar{\theta}$ plays a key role in the analysis
- Examples: the proximal gradient algorithm **[Beck and Teboulle, 2009]**, the EM algorithm **[McLachlan and Krishnan, 2007]** and variational inference **[Wainwright and Jordan, 2008]**.

MISO Algorithm

[Mairal, 2015]

Algorithm 1 MISO algorithm

Initialization: given an initial parameter estimate $\hat{\theta}^{(0)}$, for all $i \in [\![1, n]\!]$ compute a surrogate function $\vartheta \to \widehat{\mathcal{L}}_i(\hat{\theta}^{(0)}; \vartheta)$.

Iteration k: given the current estimate $\hat{\theta}^{(k)}$:

1. Pick i_k uniformly from $[\![1, n]\!]$.
2. Update $\mathcal{A}_i^{k+1}(\theta)$ as:

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

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

- Extension to Latent Data Models?
- How do those surrogates functions look like when there exists a dependence on a latent variable
- Can we derive a general algorithm?

Examples

- For smooth function $\mathcal{L}_i(\theta)$

$$\hat{\mathcal{L}}_i(\cdot, \bar{\theta}) : \theta \mapsto \mathcal{L}_i(\bar{\theta}) + \nabla \mathcal{L}_i(\bar{\theta})^\top (\theta - \bar{\theta}) + \frac{L}{2} \|\theta - \bar{\theta}\|_2^2$$

Leading to Gradient Descent Algorithm

- For nonsmooth function $\mathcal{L}_i(\theta) + \psi_i(\theta)$

$$\hat{\mathcal{L}}_i(\cdot, \bar{\theta}) : \theta \mapsto \mathcal{L}_i(\bar{\theta}) + \nabla \mathcal{L}_i(\bar{\theta})^\top (\theta - \bar{\theta}) + \frac{L}{2} \|\theta - \bar{\theta}\|_2^2 + \psi_i(\theta)$$

Leading to Proximal Gradient Algorithm

Existing Results

- **Nonconvex problems:** Almost Sure Convergence
- **Convex problems rates on** $\mathcal{L}(\theta^{(k)}) - \mathcal{L}^*$ **:**
  - $\mathcal{O}(nL/k)$ rate for convex objective
  - $\mathcal{O}((1 - \mu/(nL))^k)$ rate for strongly convex objective

Intractable Surrogates

The EM algorithm

- Complete likelihood, i.e., joint likelihood of the observations and the latent data: $f_i(z_i, y_i, \theta)$
- Likelihood of the observations: $g_i(\theta) := \int_{\mathsf{Z}} f_i(z_i, y_i, \theta)\mu_i(\mathrm{d}z_i)$
- Objective function: $\mathcal{L}_i(\theta) := -\log g_i(\theta)$

$$\mathcal{L}_i(\vartheta) = -\log g_i(\vartheta) = -\log \int_{\mathsf{Z}} f_i(z_i, y_i, \vartheta)\mu_i(\mathrm{d}z_i) = -\log \int_{\mathsf{Z}} f_i(z_i, y_i, \vartheta)\frac{p_i(z_i, \theta)}{p_i(z_i, \theta)}\mu_i(\mathrm{d}z_i)$$

Jensen Inequality →

$$\leq \int_{\mathsf{Z}} \log \frac{p_i(z_i, \vartheta)}{f_i(z_i, y_i, \theta)} p_i(z_i, \theta)\mu_i(\mathrm{d}z_i)$$

- Kullback Leibler (KL) surrogate [Neal and Hinton, 1998]

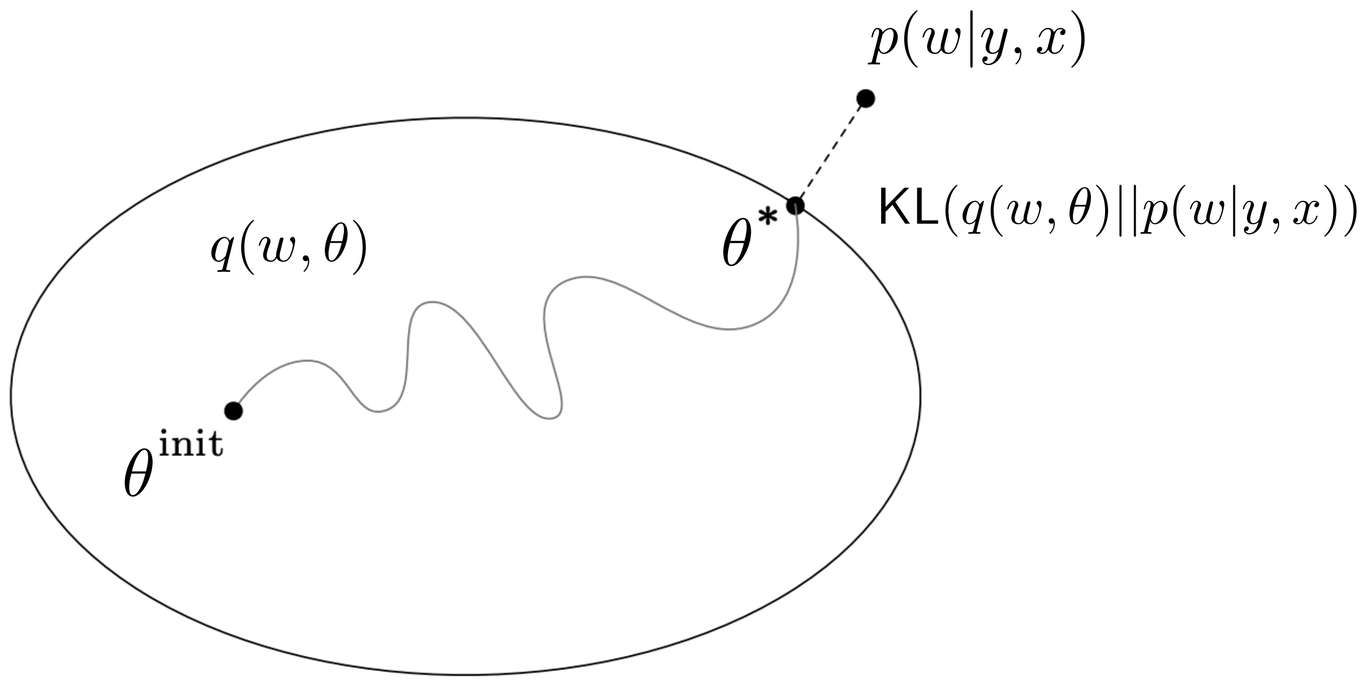
$$\hat{\mathcal{L}}_i(\theta, \vartheta) := \int_{\mathsf{Z}} \log \frac{p_i(z_i, \vartheta)}{f_i(z_i, y_i, \theta)} p_i(z_i, \theta)\mu_i(\mathrm{d}z_i) = \mathsf{KL}(p_i(z_i, \theta)||p_i(z_i, \vartheta)) + \mathcal{L}_i(\vartheta)$$

Intractable KL term

Intractable Surrogates

Variational Inference (VI)

- Input-output pairs $((x_i, y_i), 1 \leq i \leq n)$ and w a global latent variable with a prior distribution $\pi(w)$
- We want to minimize the KL between the variational candidate $q(w, \theta)$ and the true posterior $p(w|y, x)$

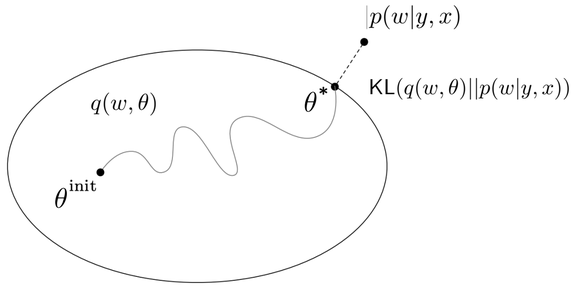


- KL term is intracable: VI optimizes the Evidence Lower Bound (ELBO)

$$\mathcal{L}(\theta) := -\mathbb{E}_{q(w;\theta)}\big[\log p(y|x, w)\big] + \mathbb{E}_{q(w;\theta)}\big[\log q(w;\theta)/\pi(w)\big]$$

- ELBO is a lower bound of the incomplete log likelihood.
- Maximizing ELBO minimizes the KL
- ‘Data term’ fits to the data and ‘KL’ term fits to the prior

Intractable Surrogates



Variational Inference (VI)

- Input-output pairs $((x_i, y_i), 1 \leq i \leq n)$ and w a global latent variable with a prior distribution $\pi(w)$
- We want to minimize the KL between the variational candidate $q(w, \theta)$ and the true posterior $p(w|y, x)$
- Individual Objective function:

$$\mathcal{L}_i(\theta) := -\mathbb{E}_{q(w;\theta)}\big[\log p(y_i|x_i, w)\big] + \frac{1}{n}\mathbb{E}_{q(w;\theta)}\big[\log q(w;\theta)/\pi(w)\big]$$

- Quadratic Surrogate function:

$$\hat{\mathcal{L}}_i(\cdot, \vartheta) : \theta \mapsto \mathcal{L}_i(\vartheta) + (\nabla r_i(\vartheta) + \nabla d_i(\vartheta))^\top(\theta - \vartheta) + \frac{L}{2}\|\theta - \vartheta\|_2^2$$

- Reparametrization trick [Blundell+, 2015]:

Let $t : \mathbb{R}^d \times \Theta \mapsto \mathbb{R}^d$ be a differentiable function w.r.t. ϑ s.t. $w = t(z, \vartheta) \sim q(\cdot, \vartheta)$ and $z \sim \mathcal{N}_p(0, I)$

$$\nabla r_i(\vartheta) = \mathbb{E}_{z\sim\mathcal{N}_p(0,I)}[\mathsf{J}_\theta^t(z, \vartheta)\nabla_w \log p(y_i|x_i, w)|_{w=t(z,\vartheta)}]$$

Intractable gradient term

MISSO Algorithm

Algorithm Formulation

Algorithm 2 MISSO algorithm

Initialization: $\hat{\theta}^{(0)}$; a sequence of non-negative numbers $\{M_{(k)}\}_{k=0}^{\infty}$.

For all $i \in [\![1, n]\!]$, draw $M_{(0)}$ samples from $p_i(\cdot; \hat{\theta}^{(0)})$ and $\widetilde{\mathcal{A}}_i^0(\theta) := \widetilde{\mathcal{L}}_i(\theta; \hat{\theta}^{(0)}, \{z_{i,m}^{(0)}\}_{m=1}^{M_{(k)}})$.

Iteration k: given the current estimate $\hat{\theta}^{(k)}$:

1. Pick a function index i_k uniformly on $[\![1, n]\!]$.
2. Draw $M_{(k)}$ Monte-Carlo samples from $p_i(\cdot; \hat{\theta}^{(k)})$.
3. Update the individual surrogate functions recursively as:

$$\widetilde{\mathcal{A}}_i^{k+1}(\theta) = \begin{cases} \widetilde{\mathcal{L}}_i(\theta; \hat{\theta}^{(k)}, \{z_{i,m}^{(k)}\}_{m=1}^{M_{(k)}}), & \text{if } i = i_k \\ \widetilde{\mathcal{A}}_i^k(\theta), & \text{otherwise.} \end{cases} \tag{13}$$

4. Set $\hat{\theta}^{(k+1)} \in \arg\min_{\theta \in \Theta} \widetilde{\mathcal{L}}^{(k+1)}(\theta) := \frac{1}{n} \sum_{i=1}^{n} \widetilde{\mathcal{A}}_i^{k+1}(\theta)$.

Class of Surrogate Functions

- Set of latent variables $(z_i \in \mathsf{Z}^m, 1 \leq i \leq n)$
- $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ Family of probability densities with respect to μ_i
- There exists function $r_i(\theta, \vartheta, z_i)$ such that:

$$\hat{\mathcal{L}}_i(\theta, \vartheta) := \int_{\mathsf{Z}} r_i(\theta, \vartheta, z_i) p_i(z_i; \vartheta) \mu_i(\mathrm{d}z_i)$$

- $\hat{\mathcal{L}}_i(\theta, \vartheta)$ fully defined by the pair $(r_i(\theta, \vartheta, z_i), p_i(z_i; \vartheta))$

Minimization by Incremental Stochastic Surrogate Optimization} (MISSO) method:

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

where $\{z_m\}_{m=1}^{M}$ is the Monte Carlo batch sampled from $p_i(z_i; \vartheta)$ either directly or using MCMC

Analysis for Constrained Optimization

- **Constrained** optimization, onsider the following **stationarity measure**:

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

where $g_+(\bar{\theta}) := \max\{0, g(\bar{\theta})\}$ and $g_-(\bar{\theta}) := -\min\{0, g(\bar{\theta})\}$ denote the **positive** and **negative** part of $g(\overline{\boldsymbol{\theta}})$,respectively.

- $\bar{\theta}$ is a stationary point if and only if $g_-(\bar{\theta}) = 0$ **[Fletcher+, 2002]**.

- Furthermore, suppose that the sequence $\{\theta^{(k)}\}_{k \geq 0}$ has a limit point $\bar{\theta}$ that is a stationary point, then one has:

$$\lim_{k \to \infty} g_-(\theta^{(k)}) = 0$$

Global Convergence

Assumptions

(S1) Upper bounding surrogate $\hat{\mathcal{L}}_i(\theta, \vartheta) \geq \mathcal{L}_i(\theta)$ with **equality** if $\theta = \vartheta$

(S2) The **approximation error** $\widehat{e}(\theta; \{\vartheta_i\}_{i=1}^n) := \dfrac{1}{n}\sum_{i=1}^{n} \hat{\mathcal{L}}_i(\theta, \vartheta_i) - \mathcal{L}(\theta)$

is defined on $\Theta\epsilon = \{\theta \in \mathbb{R}^d, \inf_{\theta' \in \Theta} \|\theta - \theta'\| < \epsilon\}$ and satisfies: $\left\|\nabla \widehat{e}(\theta; \{\vartheta_i\}_{i=1}^n)\right\|^2 \leq 2L\, \widehat{e}(\theta; \{\vartheta_i\}_{i=1}^n)$

(H1) $r_i(\theta, \vartheta, z_i)$ is **convex** and **lower bounded**

(H2) There exists the following **constants**:

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

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

Global Convergence

Non-Asymptotic Analysis

Theorem

Under **(S1)**, **(S2)**, **(H1)**, **(H2)** and define the following quantity:

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

Then we have the following bounds:

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

$$\mathbb{E}\left[g_-\left(\hat{\theta}^{(K)}\right)\right] \leq \sqrt{\frac{\Delta_{(K_{\max})}}{K_{\max}}} + \frac{C_{\max}}{K_{\max}} \sum_{k=0}^{K_{\max}-1} M_{(k)}^{-1/2}$$

Asymptotic Analysis

Theorem

Also, assume $\sum_{k=0}^{\infty} M_{(k)}^{-1/2} < \infty$ (non decreasing sequence)

- $\lim_{k\to\infty} g_-\left(\hat{\theta}^{(k)}\right) = 0$
- $\lim_{k\to\infty} \mathcal{L}\left(\hat{\theta}^{(k)}\right) = \mathcal{L}^*$

Remarks

- $\Delta_{(K_{\max})}$ **is finite for any** $K_{\max} \in \mathbb{N}$
- **MISO as a special case of MISSO**

$$C_{\mathsf{r}} = C_{\mathsf{gr}} = 0$$

- **Non-asymptotic rate of**

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

- **MISSO sequence.** $\{\theta^{(k)}\}_{k\geq 0}$ **satisfies an** ***asymptotic stationary*** **point condition**

Numerical Applications

Logistic Regression with Missing Covariates

- $y = (y_i, 1 \leq i \leq n)$ vector of binary responses and $z_i = (z_{i,p}) \in \mathbb{R}^d$ covariates
- z_i is not fully observed:
  - $z_{i,mis}$ missing values and $z_{i,obs}$ observed
- $z = (z_i, 1 \leq i \leq n) \sim \mathcal{N}(\beta, \Omega)$ where $\beta \in \mathbb{R}^d$
- Finding the structure of the missing data
- Logit model

$$\text{logit}(\mathbb{P}(y_{ij} = 0 | z_i)) = d_{ij}^\top z_i$$

- Exponential family with statistics $\tilde{S}_i(z_i) := (z_i, z_i^\top z_i)$

MISSO Update

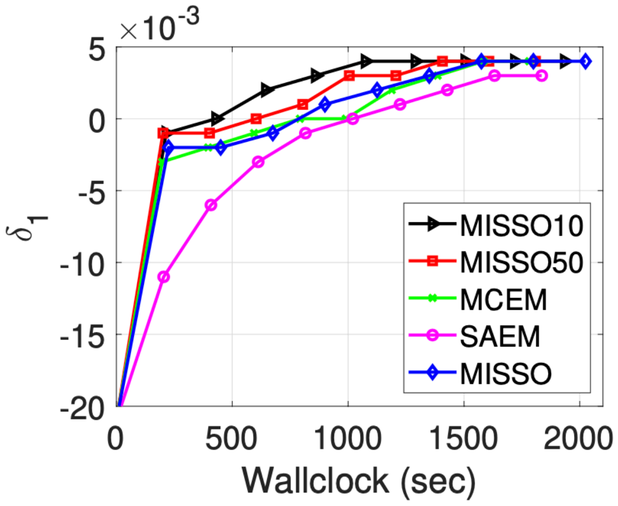
- Pick a set I_k, sample a Monte Carlo batch and update the statistics as follows:

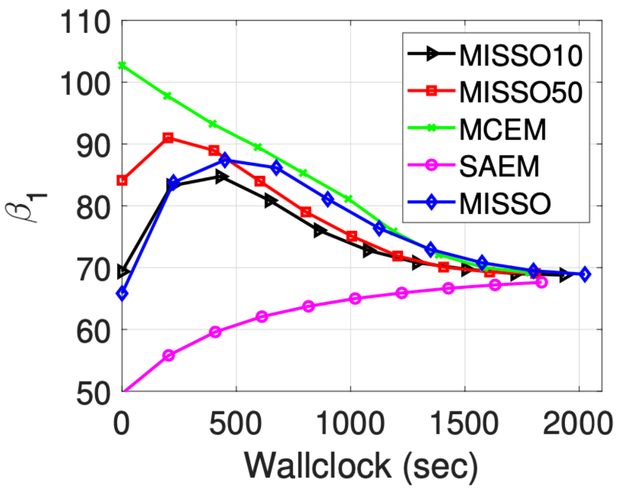
$$(s_i^{1,k}, s_i^{2,k}) = \begin{cases} \left(\frac{1}{M_k} \sum_{m=0}^{M_k - 1} z_i^{k,m}, \frac{1}{M_k} \sum_{m=0}^{M_k - 1} (z_i^{k,m})^\top z_i^{k,m} \right) & \text{if } i \in I_k \\ (s_i^{1,k-1}, s_i^{2,k-1}) & \text{otherwise} \end{cases}$$

- Then $\beta^k = \frac{1}{n} \sum_{i=1}^n s_i^{1,k} \quad \Omega^k = \frac{1}{n} \sum_{i=1}^n s_i^{2,k} - (\beta^k)^\top \beta^k$

Experiments - TraumaBase dataset

- TraumaBase (http://traumabase.eu) dataset:
  - **15 trauma centers** in France
  - Measurements from the initial to last stage of trauma.
  - **6384** patients and **d=16** quantitative variables
- Predict binary response: severe trauma or not.





Numerical Applications

Fitting Bayesian LeNet5 on MNIST and Bayesian ResNet18 on CIFAR

- **Weight prior:** $p(w) = \mathcal{N}(0, I)$ and $p(y_i|x_i, w) = \mathrm{Softmax}(f(x_i, w))$ where f is a NN
- **Variational candidate:** for any layer: $q(w_\ell, \theta_\ell)$ is a Gaussian distribution $\mathcal{N}(\mu_\ell, \sigma^2 I)$

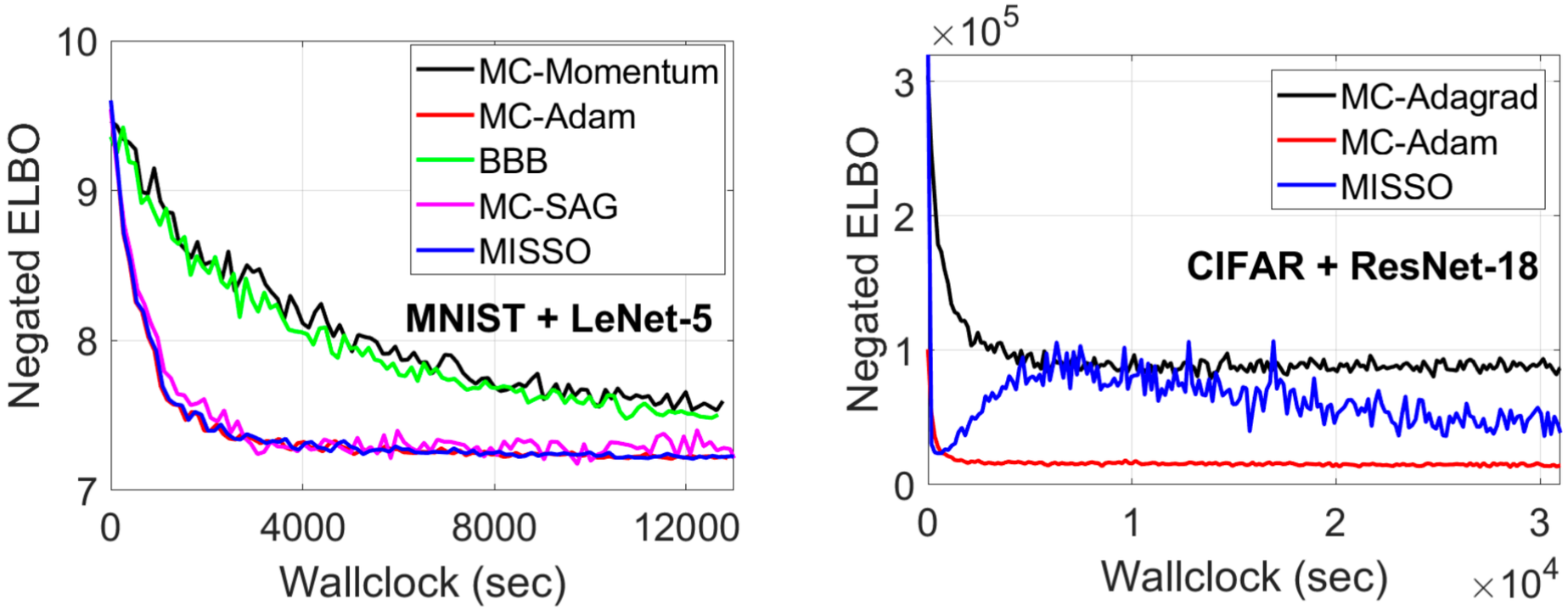


Figure 2: Negated ELBO versus time elapsed for fitting (Left) Bayesian LeNet-5 on MNIST and (Right) Bayesian ResNet-18 on CIFAR-10. The solid curve is obtained from averaging over 5 independent runs of the methods. The convergence is plotted against the wallclock time.

Conclusion

Take-Away

- We derived an **incremental** method for the optimization problem in machine learning.
  - When the objective function is a likelihood or not
  - For latent data models

- We conducted **finite-time analysis** of these methods for **nonconvex** loss functions and **non necessarily gradient** methods.

Perspective

- Incremental algorithms: choice of the indices at each iteration.
  - **Optimal sampling strategies**: **[Le Roux+, 2012]** or **[Horvath and Richatrik, 2018]**.
  - **Optimal mini-batch size** of stochastic and incremental algorithms. See **[Gower+, 2019]** (variance-cost trade off).

- **Interplay** between the Monte Carlo batch and the mini-batch of indices drawn at each iteration (**bias-variance trade off**).

Thank You !