MISSO: Minimization by Incremental Stochastic Surrogate for large-scale nonconvex Optimization

PGMODays 2018

Belhal Karimi and Eric Moulines

Nov. 21^{th}, 2018

Outline

1) Motivation: Large-scale machine learning

2) Majorization-minorization principle

3) Incremental Surrogate minimization scheme

4) MISSO: Stochastic incremental framework

5) Application to variational inference

Large-scale machine learning

Constrained Minimization of large sum of functions

We are interested in the minimization of a large finite-sum of functions:

$$\min_{\theta \in \Theta} \left[f(\theta) \triangleq \sum_{i=1}^{N} f_i(\theta) \right] \tag{1}$$

where Θ is a **convex subset** of $\mathbb{R}^p$, for all $i \in [\![N]\!]$, $f_i : \mathbb{R}^p \to \mathbb{R}$ are continuously differentiable, **bounded from below** and possibly **nonconvex**.

Some examples Given data points $(x_i, i \in [\![N]\!])$ and observations $(y_i, i \in [\![N]\!])$

- Maximum likelihood estimation: $f(\theta) \triangleq -\sum_{i=1}^{N} \log p_i(y_i, \theta)$
- Variational inference: $f(\theta) \triangleq \sum_{i=1}^{N} \mathrm{KL}\left(q_i(w;\theta) || p_i(w|y_i, x_i)\right)$
- Logistic regression: $f(\theta) \triangleq \sum_{i=1}^{N} \log(1 + e^{-y_i <\theta, x_i>})$

Majorization-Minimization principle

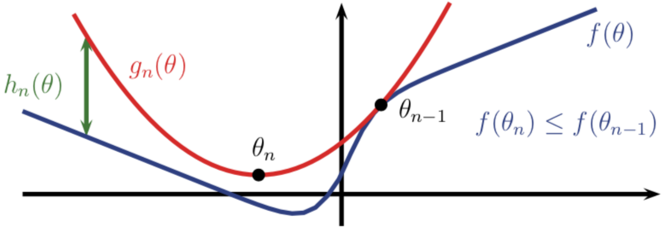

Figure 1: MM principle. Plot from [Mairal, 2015]

- Iteratively minimize locally tight upper bounds on the objective
- Drives the objective function downwards
- Examples: the proximal gradient algorithm (Beck and Teboulle, 2009), the EM algorithm (McLachlan and Krishnan, 2007) and variational inference (Wainwright and Jordan, 2008).

Notations and Assumptions

- Constrained optimization: Θ convex subset of $\mathbb{R}^p$. And $\mathcal{T}(\Theta)$ neighborhood of Θ.
- For all $i \in [\![N]\!]$, f_i is continuously differentiable on $\mathcal{T}(\Theta)$.
- For all $i \in [\![N]\!]$, f_i is bounded from below, i.e. there exist a constant $M_i \in \mathbb{R}$ such as for all $\theta \in \Theta$, $f_i(\theta) \geq M_i$.
- $f_{i,\theta} : \mathbb{R}^p \to \mathbb{R}$ is a surrogate of f_i at θ if the following properties are satisfied:
  1. the function $\vartheta \to f_{i,\theta}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$
  2. for all $\vartheta \in \Theta$, $f_{i,\theta}(\vartheta) \geq f_i(\vartheta)$, $f_{i,\theta}(\theta) = f_i(\theta)$ and $\nabla f_{i,\theta}(\vartheta)\Big|_{\vartheta=\theta} = \nabla f_i(\vartheta)\Big|_{\vartheta=\theta}$.
- A sequence $(\theta^k)_{k\geq 0}$ satisfies the asymptotic stationary point condition if

$$\liminf_{k\to\infty} \inf_{\theta\in\Theta} \frac{\langle \nabla f(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq 0. \qquad (2)$$

Incremental Surrogate Minimization

The incremental scheme of (Mairal, 2015) computes surrogate functions, at each iteration of the algorithm, for a mini-batch of components:

Algorithm 1 MISO algorithm

Initialization: given an initial parameter estimate θ^0, for all $i \in [\![N]\!]$ compute a surrogate function $\vartheta \to f_{i,\theta^0}(\vartheta)$.

Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![N]\!], \mathrm{card}(A) = p\}$
2. For all $i \in I_k$ compute $\vartheta \to f_{i,\theta^{k-1}}(\vartheta)$, a surrogate of f_i at θ^{k-1}.
3. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N a_i^k(\vartheta)$ where $a_i^k(\vartheta)$ are defined recursively as follows:

$$a_i^k(\vartheta) \triangleq \begin{cases} f_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ a_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{3}$$

Intractable surrogate functions

- In many cases of interest those surrogates are intractable.
- Denote by $z = (z_i \in \mathsf{Z}_i, i \in [\![N]\!]) \in \mathsf{Z}$ where $\mathsf{Z} = \bigtimes_{i=1}^{N} \mathsf{Z}_i$ where Z_i is a subset of $\mathbb{R}^{m_i}$ as set of latent variables.
- For all $i \in [\![N]\!]$, let μ_i be a σ-finite measure on the Borel σ-algebra $\mathcal{Z}_i = \mathcal{B}(\mathsf{Z}_i)$.
- $\mathcal{P}_i = \{p_i(z_i, \theta); \theta \in \Theta\}$ be a family of probability densities with respect to μ_i, and $r_{i,\theta} : \mathsf{Z}_i \times \Theta \to \mathbb{R}$ be functions such that:

$$\boxed{f_{i,\theta}(\vartheta) \triangleq \int_{\mathsf{Z}_i} r_{i,\theta}(z_i, \vartheta) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \quad \text{for all } (\theta, \vartheta) \in \Theta^2.} \tag{4}$$

The surrogate function denoted $f_{i,\theta}(\vartheta)$ is fully defined by the pair $(r_{i,\theta}(z_i, \vartheta), p_i(z_i, \theta))$.

Examples of intractable surrogates

Incremental EM algorithm

- In the missing data context, let $c_i(z_i, \theta)$ be the joint likelihood of the observations and the latent data referred to as the complete likelihood.
- $g_i(\theta) \triangleq \int_{\mathsf{Z}_i} c_i(z_i, \theta)\mu_i(\mathrm{d}z_i)$ is the likelihood of the observations (in which the latent variables are marginalized).

The incremental EM algorithm falls into the incremental MM framework:

- For $i \in [\![N]\!]$ and $\theta \in \Theta$ the loss function $f_i(\theta) \triangleq -\log g_i(\theta)$
- for $\vartheta \in \Theta$ the surrogate function $f_{i,\theta}(\vartheta)$, introduced in the pioneering paper (Neal and Hinton, 1998), is defined by

$$f_{i,\theta}(\vartheta) \triangleq \int_{\mathsf{Z}_i} \log \frac{p_i(z_i, \theta)}{c_i(z_i, \vartheta)} p_i(z_i, \theta)\mu_i(\mathrm{d}z_i) = \mathrm{KL}(p_i(z_i, \theta) \parallel p_i(z_i, \vartheta)) + f_i(\vartheta) \tag{5}$$

In most cases, this surrogate is intractable.

Examples of intractable surrogates

Variational Inference

- Let $x = (x_i, i \in [\![N]\!])$ and $y = (y_i, i \in [\![N]\!])$ be i.i.d. input-output pairs and w be a global latent variable taking values in W a subset of $\mathbb{R}^J$.
- A natural decomposition of the joint distribution is:

$$p(y, x, w) = p(w) \prod_{i=1}^{N} p_i(y_i | x_i, w) \tag{6}$$

- The variational inference problem boils down to minimizing the following KL divergence:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{KL}(q(w; \theta) \parallel p(w|y, x)) = \arg\min_{\theta \in \Theta} f(\theta) \tag{7}$$

where for all $\theta \in \Theta$, $f(\theta) = \sum_{i=1}^{N} f_i(\theta)$ with :

$$f_i(\theta) \triangleq -\int_{\mathsf{W}} q(w; \theta) \log p_i(y_i, x_i | w) \mathrm{d}w + \frac{1}{N} \mathrm{KL}(q(w; \theta) \parallel p(w)) \tag{8}$$

Examples of intractable surrogates

Variational Inference

- Does not scale to large data since evaluating the reconstruction term in (8) requires calculations over the entire dataset.
- Optimization using the incremental framework from (Mairal, 2013) as in (Hoffman et al., 2013; Titsias and Lázaro-Gredilla, 2014; Kucukelbir et al., 2017; Kingma and Welling, 2013).
- Quadratic surrogate at $\theta \in \Theta$:

$$f_{i,\theta}(\vartheta) \triangleq f_i(\theta) + \nabla f_i(\theta)^\top (\vartheta - \theta) + \frac{L}{2}\|\vartheta - \theta\|_2^2 \tag{9}$$

where $\|\cdot\|_2$ is the ℓ_2-norm and L is an upper bound of the spectral norm of the Hessian of f_i at θ.

The reconstruction integral term can not be calculated in complex models such as Bayesian neural networks (Neal, 2012; Gal, 2016).

MISSO: Stochastic incremental scheme

We propose an incremental stochastic surrogate scheme called MISSO (Minimization by Incremental Stochastic Surrogate Optimization):

- For $i \in [\![N]\!]$, $\hat{f}_{i,\theta}(\vartheta)$ is a Monte Carlo approximation of $f_{i,\theta}(\vartheta)$:

$$\hat{f}_{i,\theta}(\vartheta) \triangleq \frac{1}{M} \sum_{m=0}^{M-1} r_{i,\theta}(z_i^m, \vartheta) \quad \text{for all } (\theta, \vartheta) \in \Theta^2 \tag{10}$$

- $\{z_i^m\}_{m=0}^{M-1}$ is a Monte Carlo batch.
- In many cases, sampling from $p_i(z_i, \theta)$ is not an option.
- Sampled by Monte Carlo Markov Chain (MCMC) algorithm.

MISSO: Stochastic incremental scheme

Algorithm

Algorithm 2 MISSO algorithm

Initialization: given an initial parameter estimate θ^0, for all $i \in [\![N]\!]$ compute the function $\vartheta \to \hat{f}_{i,\theta^0}(\vartheta)$ defined by (10).

Iteration k: given the current estimate θ^{k-1}:

1. Pick a set I_k uniformly on $\{A \subset [\![N]\!], \mathrm{card}(A) = p\}$
2. For all $i \in I_k$, sample a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ from $p_i(z_i, \theta^{k-1})$.
3. For all $i \in I_k$, compute the function $\vartheta \to \hat{f}_{i,\theta^{k-1}}(\vartheta)$ defined by (10).
4. Set $\theta^k \in \arg\min_{\vartheta \in \Theta} \sum_{i=1}^N \hat{a}_i^k(\vartheta)$ where $\hat{a}_i^k(\vartheta)$ are defined recursively as follows:

$$\hat{a}_i^k(\vartheta) \triangleq \begin{cases} \hat{f}_{i,\theta^{k-1}}(\vartheta) & \text{if } i \in I_k \\ \hat{a}_i^{k-1}(\vartheta) & \text{otherwise} \end{cases} \tag{11}$$

MISSO: Stochastic incremental scheme

Need to control the supremum norm of **the fluctuations of the Monte Carlo approximation**.

Let $i \in [\![N]\!]$, $\{j_i(z_i, \vartheta), z_i \in \mathsf{Z}_i, \vartheta \in \Theta\}$ be a family of measurable functions, λ_i a probability measure on $\mathsf{Z}_i \times \mathcal{Z}_i$.

We define:

$$C_i(j_i) \triangleq \sup_{\theta \in \Theta} \mathbb{E}_{i,\theta}\left[\sup_{\vartheta \in \Theta}\left|\sum_{m=0}^{M-1}\left\{j_i(z_i^m, \vartheta) - \int_{\mathsf{Z}_i} j_i(z_i, \vartheta) p_i(z_i, \theta) \lambda_i(\mathrm{d}z_i)\right\}\right|\right] \tag{12}$$

$\mathbb{E}_{i,\theta}$ the expectation of the Markov chain $\{z_i^m\}_{m=0}^{\infty}$ with transition kernel $P_{i,\theta}$ and stationary distribution $p_i(z_i, \theta) \cdot \lambda_i$

In most examples, the Markov kernel $P_{i,\theta}$ is derived from an MCMC algorithm.

MISSO: Main Result

Assumptions:

- For $i \in [\![N]\!]$, $\lim_{k\to\infty} C_i(r_{i,\theta}) < \infty \quad \text{and} \quad \lim_{k\to\infty} C_i(\nabla r_{i,\theta}) < \infty$.
- $\{M_k\}_{k\geq 0}$ is a non deacreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$.

Theorem: Asymptotic results

Given the assumptions above. Let $\left(\theta^k\right)_{k\geq 1}$ be a sequence generated from $\theta^0 \in \Theta$ by the iterative application described by Algorithm 2. Then:

- $\left(f(\theta^k)\right)_{k\geq 1}$ converges almost surely.
- $\left(\theta^k\right)_{k\geq 1}$ satisfies the Asymptotic Stationary Point Condition.

Application to EM algorithm

With our notations, we define the Monte Carlo approximation of the intractable surrogate as:

$$\hat{f}_{i,\theta}(\vartheta) \triangleq \frac{1}{M} \sum_{m=0}^{M-1} \log \frac{p_i(z_i^m, \theta)}{c_i(z_i^m, \vartheta)} \quad \text{for all } i \in [\![N]\!] \text{ and } (\theta, \vartheta) \in \Theta^2. \quad (13)$$

where $\{z_i^m\}_{m=0}^{M-1}$ is a Monte Carlo batch sampled from $p_i(z_i, \theta)$ using an MCMC procedure.

The MISSO algorithm yields, at iteration k, the following update of the parameter:

$$\boxed{\theta^k \in \arg\min_{\vartheta \in \Theta} - \sum_{i=0}^{N} \frac{1}{M_{\tau_{i,k}}} \sum_{m=0}^{M_{\tau_{i,k}}-1} \log c_i(z_i^{\tau_{i,k}+1,m}, \vartheta)} \quad (14)$$

where $\{z_i^{\tau_{i,k}+1,m}\}_{m=0}^{M-1}$ is a Monte Carlo batch sampled from $p_i(z_i, \theta^{\tau_{i,k}})$ and $\tau_{i,k} = k-1$ if $i \in I_k$ and $\tau_{i,k-1}$ otherwise.

Application to Variational Inference

- Recall, the intractable surrogate:

$$f_{i,\theta}(\vartheta) \triangleq f_i(\theta) + \nabla f_i(\theta)^\top(\vartheta - \theta) + \frac{L}{2}\|\vartheta - \theta\|_2^2 \tag{15}$$

- Reparametrization trick suggested in (Kingma and Welling, 2013; Blundell et al., 2015). For $\theta \in \Theta$ and $e \in \mathbb{R}^d$, $W = t(\theta, e)$ where $e \sim \mathcal{N}_d(0, \mathrm{Id})$ has a density $q(\cdot, \theta)$. (Blundell et al., 2015, Proposition 1):

$$\nabla \int_{\mathsf{W}} \log p_i(y_i, x_i|w) q(w, \theta)\mathrm{d}w = \int_{\mathsf{W}} \mathsf{J}(\theta, e)\nabla \log p_i(y_i, x_i|t(\theta, e))\phi(e)\mathrm{d}e$$

where $\mathsf{J}(\theta, e)$ is the Jacobian of the function $t(\cdot, e)$.

- Setting $\theta^k = \frac{1}{N}\sum_{i=1}^N \theta^{\tau_{i,k}} - \gamma \sum_{i=1}^N \hat{m}_i^k$ where $\hat{m}_i^k$ are defined recursively as follows:

$$\hat{m}_i^k \triangleq \begin{cases} -\frac{1}{M_k}\sum_{m=0}^{M_k-1} \mathsf{J}(\theta, e^{k,m})\nabla_\theta \log p_i(y_i, x_i|t(\theta, e^{k,m})) + \nabla d(\theta^{k-1}) & i \in I_k \\ \hat{m}_i^{k-1} & i \notin I_k \end{cases} \tag{16}$$

Numerical Application:Training a BNN

- 2-layer Bayesian neural network on the MNIST dataset (LeCun and Cortes, 2010)
- $N = 60\,000$ handwritten digits, 28×28 images, $d = 784$
- input layer with $d = 784$ units
- a single hidden layer of $p = 100$ hyperbolic tangent units
- softmax output layer with $K = 10$ classes.
- $p(w) = \mathcal{N}(0, \mathrm{Id})$, $p(y_i|x_i, w) = \mathrm{Softmax}(f(x_i, w))$ where f is the two layers model.
- Variational distribution for layier j: $q(w_j, \theta_j)$ is a multivariate Gaussian distribution $\mathcal{N}(\rho_j, \sigma_j^2 \, \mathrm{Id})$

Numerical Application:Training a BNN

- Comparison with state-of-the-art optimizers: SAG, ADAM, Momentum and SGD.
- 2 batch sizes: 1% and 10%
- Constant learning rate of 10^{-5}

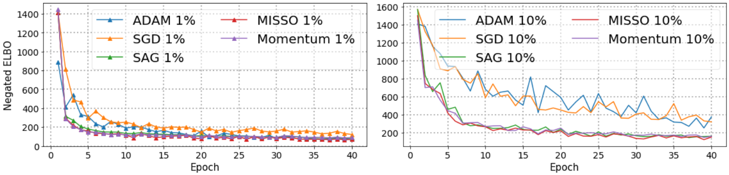

Figure 2: (Incremental Variational Inference) Convergence of the negated ELBO for 40 epochs over the training set. Runs for two different mini-batch sizes (1% left and 10% right).

Conclusion

- Unifying framework for minimization by incremental surrogate optimization with MC approximation of the surrogates.
- Covers a large class of nonconvex optimization algorithms used in machine learning.
- The incremental approach reduces significantly the variance. (see SAG, MISO)

Future works include:

- Non asymptotic convergence results for both convex and nonconvex objective functions
- Fixed Monte Carlo batch size convergence guarantees

B.K. and E. Moulines, *MISSO: Minimization by Incremental Stochastic Surrogate for large-scale nonconvex Optimization*, 2018

B.K., M. Lavielle and E. Moulines, *On the Convergence Properties of the Mini-Batch EM and MCEM algorithms*, 2018

Thank you!

References

A. Beck and M. Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci*, 2:pp. 183–202, 2009.

C. Blundell, J. Cornebise, K. Kavukcuoglu, and D. Wierstra. Weight uncertainty in neural networks. In *Proceedings of the 32Nd International Conference on International Conference on Machine Learning - Volume 37*, ICML'15, pages 1613–1622. JMLR.org, 2015. URL `http://dl.acm.org/citation.cfm?id=3045118.3045290`.

Y. Gal. *Uncertainty in Deep Learning*. PhD thesis, University of Cambridge, 2016.

Blei D.M. Hoffman, M. D., C. Wang, and J.W. Paisley. Stochastic variational inference. *Journal of Machine Learning Research*, 14(1): 1303–1347, 2013.

D.P. Kingma and M. Welling. Auto-encoding variational bayes. *CoRR*, abs/1312.6114, 2013. URL http://dblp.uni-trier.de/db/journals/corr/corr1312.html#KingmaW13.

A. Kucukelbir, D. Tran, R. Ranganath, A. Gelman, and D. M. Blei. Automatic differentiation variational inference. *J. Mach. Learn. Res.*, 18(1):430–474, January 2017. ISSN 1532-4435. URL http://dl.acm.org/citation.cfm?id=3122009.3122023.

Y. LeCun and C. Cortes. MNIST handwritten digit database. 2010. URL http://yann.lecun.com/exdb/mnist/.

Julien Mairal. Stochastic majorization-minimization algorithms for large-scale optimization. In *NIPS*, pages 2283–2291, 2013.

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

G. McLachlan and T. Krishnan. The EM Algorithm and Extensions. 2007.

R. Neal and G.E. Hinton. A view of the em algorithm that justifies incremental, sparse, and other variants. In *Learning in Graphical Models*, pages 355–368. Kluwer Academic Publishers, 1998.

R. M. Neal. Bayesian learning for neural networks. *Springer Science Business Media*, 118, 2012.

J.W. Paisley, D.M. Blei, and M.I. Jordan. Variational bayesian inference with stochastic search. In *ICML*. icml.cc / Omnipress, 2012. URL `http://dblp.uni-trier.de/db/conf/icml/icml2012.html#PaisleyBJ12`.

M. Titsias and M. Lázaro-Gredilla. Doubly stochastic variational bayes for non-conjugate inference. In Eric P. Xing and Tony Jebara, editors, *Proceedings of the 31st International Conference on Machine Learning*, volume 32 of *Proceedings of Machine Learning Research*, pages 1971–1979, Bejing, China, 22–24 Jun 2014. PMLR. URL `http://proceedings.mlr.press/v32/titsias14.html`.

M. J. Wainwright and M. I. Jordan. Graphical models, exponential families, and variational inference. *Found. Trends Mach. Learn.*, 1:pp. 1–305, 2008.

Appendix

Logistic Regression

- $y = (y_i, i \in [\![N]\!])$ vector of binary responses and $z_i = (z_{i,p}) \in \mathbb{R}^p$ vector latent data independent and marginally distributed according to $\mathcal{N}(\beta, \Omega)$
- Model defined by

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

- Model belongs to the curved exponential family. Sufficient statistics are $\tilde{S}_i(z_i) \triangleq (z_i, z_i^\top z_i)$.
- MISSO algorithm consists in picking a set I_k, sampling a Monte Carlo batch $\{z_i^{k,m}\}_{m=0}^{M_k-1}$ for $i \in I_k$ and computing the quantities $(s_i^{1,k}, s_i^{2,k})$ as follows:

$$(s_i^{1,k}, s_i^{2,k}) = \begin{cases} \left(\frac{1}{M_k}\sum_{m=0}^{M_k-1} z_i^{k,m}, \frac{1}{M_k}\sum_{m=0}^{M_k-1} (z_i^{k,m})^\top z_i^{k,m}\right) & \text{if } i \in I_k \\ (s_i^{1,k-1}, s_i^{2,k-1}) & \text{otherwise} \end{cases} \tag{17}$$

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

Logistic Regression

- $p = 3$, $N = 1200$ and for all $i \in [\![N]\!]$, $n_i = 15$.
- $d_{ij,1} = 1$, $d_{ij,2} = -20 + (j-1) * 5$ and $d_{ij,3} = 10\lceil 3i/N \rceil$.
- Data generating values:
  $(\beta_1 = -4, \beta_2 = -0.5, \beta_3 = 1, \omega_1 = 0.3, \omega_2 = 0.2, \omega_3 = 0.2)$.
- $M_k \triangleq M_0 + k^2$ with $M_0 = 50$.

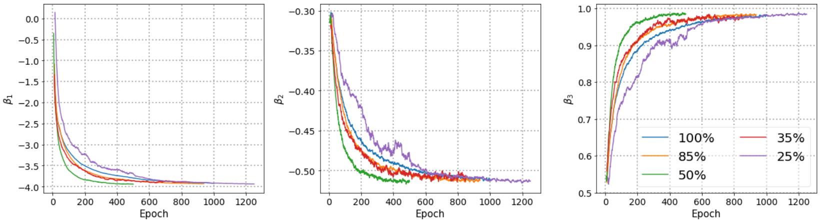

Figure 3: (Incremental MCEM) Convergence of the vector of fixed parameters β for different batch sizes function of passes over the data.