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

Belhal Karimi
CMAP, Ecole Polytechnique, FR
belhal.karimi@polytechnique.edu

Eric Moulines
CMAP, Ecole Polytechnique, FR
eric.moulines@polytechnique.edu

1 Introduction

We are interested in the constrained minimization of a large sum of nonconvex functions defined as $\min_{\theta\in\Theta}\left[f(\theta) \triangleq \sum_{i=1}^{N} f_i(\theta)\right]$ where Θ is a convex subset of $\mathbb{R}^p$ and or all $i \in [\![N]\!]$, $f_i : \mathbb{R}^p \to \mathbb{R}$ are continuously differentiable, bounded from below and possibly nonconvex. In this paper, we solve this minimization problem using an MM algorithm [Lange, 2016, Razaviyayn et al., 2013] which works by finding iteratively a surrogate function that majorizes the objective function. MM algorithms are very popular in machine learning and computational statistics [Lange, 2016]. Examples include proximal gradient algorithms [Parikh and Boyd, 2014, Mishchenko et al., 2018], the Expectation-Maximization (EM) algorithm [McLachlan and Krishnan, 2007] and some variational inference methods [Wainwright and Jordan, 2008]. When the objective function is a finite-sum, [Mairal, 2015] developed an incremental MM scheme, called MISO, taking advantage of the finite-sum structure with a cost per iteration that is independent of N. However, the MISO framework rests upon the computation of tractable surrogates such as quadratic functions. Yet, in many cases, those surrogates are intractable and need to be approximated. This is the case in particular in Bayesian Deep Learning [Ghahramani, 2015, Ranganath et al., 2014, Kingma and Welling, 2013]. Ultimately, MISO convergence guarantees can not be applied on those cases where approximation of surrogates are used; they often rely on Robbins and Monro [Robbins and Monro, 1951] convergence results for stochastic optimization.

In this contribution, we propose an incremental MM algorithm, called MISSO (Minimization by Incremental Stochastic Surrogate Optimization) when the natural surrogate functions are intractable and should be approximated, for example by Monte Carlo integration, and provide asymptotic convergence guarantees.

2 Minimization by Incremental Stochastic Surrogate Optimization (MISSO)

M 1. *For $i \in [\![N]\!]$, f_i is continuously differentiable on a neighborhood $\mathcal{T}(\Theta)$ of Θ and is bounded from below*

For any $\theta \in \Theta$ and $i \in [\![N]\!]$, we say, following [Mairal, 2015] that a function $f_{i,\theta} : \mathbb{R}^p \to \mathbb{R}$ is a surrogate of f_i at θ if the function $\vartheta \to f_{i,\theta}(\vartheta)$ is continuously differentiable on $\mathcal{T}(\Theta)$, $f_{i,\theta}(\vartheta) \geq f_i(\vartheta)$, $f_{i,\theta}(\theta) = f_i(\theta)$ and $\nabla f_{i,\theta}(\vartheta)\Big|_{\vartheta=\theta} = \nabla f_i(\vartheta)\Big|_{\vartheta=\theta}$.

M 2. *For $i \in [\![N]\!]$, $h_{i,\theta} \triangleq f_{i,\theta} - f_i$ is L-smooth, i.e. for all $(\theta, \vartheta, \vartheta') \in \Theta^3$: $|\nabla h_{i,\theta}(\vartheta) - \nabla h_{i,\theta}(\vartheta')| \leq L|\vartheta - \vartheta'|$*

We introduce an incremental scheme to deal with the case of intractable surrogate functions. We assume that the surrogate can be expressed as an integral over a set of latent variables, denoted $z = (z_i \in \mathsf{Z}_i, i \in [\![N]\!]) \in \mathsf{Z}$ where $\mathsf{Z} = \bigtimes_{i=1}^{N} \mathsf{Z}_i$ with Z_i a subset of $\mathbb{R}^{m_i}$. For all $i \in [\![N]\!]$, let μ_i

Third workshop on Bayesian Deep Learning (NeurIPS 2018), Montréal, Canada.

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:

$$f_{i,\theta}(\vartheta) \triangleq \int_{\mathsf{Z}_i} r_{i,\theta}(z_i, \vartheta) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \quad \text{for all } (\theta, \vartheta) \in \Theta^2. \tag{1}$$

Our scheme is based on the computation, at each iteration, of stochastic surrogate functions for a mini-batch of components. For $i \in [\![N]\!]$, the stochastic surrogate function, noted $\hat{f}_{i,\theta}^M(\vartheta)$ is a Monte Carlo approximation of the surrogate function $f_{i,\theta}(\vartheta)$ defined by (1) such that:

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

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

Algorithm 1 MISSO algorithm

Initialization: given an initial estimate θ^0, compute $\vartheta \to \hat{f}_{i,\theta^0}^{M_0}(\vartheta)$ defined by (2) for $i \in [\![N]\!]$.

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 $p_i(z_i, \theta^{k-1})$ and compute $\vartheta \to \hat{f}_{i,\theta^{k-1}}^{M_k}(\vartheta)$ defined by (2)
3. 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: $\hat{a}_i^k(\vartheta) \triangleq \hat{f}_{i,\theta^{k-1}}^{M_k}(\vartheta)$ if $i \in I_k$ and $\hat{a}_i^{k-1}(\vartheta)$ otherwise

Whether we use MCMC or direct simulation, we 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. We define:

$$C_i(j_i) \triangleq \sup_{\theta \in \Theta} \sup_{M > 0} M^{-1/2} \mathbb{E}_{i,\theta}^M \left[\sup_{\vartheta \in \Theta} \left| \sum_{m=0}^{M-1} \left\{ j_i(z_i^m, \vartheta) - \int_{\mathsf{Z}_i} j_i(z_i, \vartheta) p_i(z_i, \theta) \mu_i(\mathrm{d}z_i) \right\} \right| \right] \tag{3}$$

where we denote by $\mathbb{E}_{i,\theta}^M$ is the expectation of the samples $\{z_i^m\}_{m=0}^{M-1}$.

M 3. *For all $i \in [\![N]\!]$ and $\theta \in \Theta$: $\lim_{k\to\infty} C_i(r_{i,\theta}) < \infty$ and $\lim_{k\to\infty} C_i(\nabla r_{i,\theta}) < \infty$.*

M 4. *$\{M_k\}_{k\geq 0}$ is a non deacreasing sequence of integers which satisfies $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$.*

Theorem 1. *Assume* **M1-M4**. *Let $\left(\theta^k\right)_{k\geq 1}$ be a sequence generated from $\theta^0 \in \Theta$ by the iterative application described by Algorithm 1. Then:*

(i) $\left(f(\theta^k)\right)_{k\geq 1}$ converges almost surely and $\left(\theta^k\right)_{k\geq 1}$ satisfies the Asymptotic Stationary Point Condition of [Mairal, 2015].

Proof. The proof is postponed to the supplementary material. □

3 Incremental Variational Inference for Bayesian Deep Learning

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 $p(y, x, w) = p(w) \prod_{i=1}^N p_i(y_i|x_i, w)$. The goal is to calculate the posterior distribution $p(w|y, x)$. The classical variational inference problem boils down to minimizing the following Kullback Leibler (KL) divergence:

$$\theta^* = \arg\min_{\theta \in \Theta} \mathrm{KL}(q(w; \theta) \parallel p(w|y, x)) = \arg\min_{\theta \in \Theta} f(\theta) = \arg\min_{\theta \in \Theta} \sum_{i=1}^N f_i(\theta) \tag{4}$$

where $q(w;\theta)$ belongs to the multivariate Gaussian family and f_i is defined as:

$$f_i(\theta) \triangleq -\int_{\mathsf{W}} \log p_i(y_i|x_i,w)q(w;\theta)\mathrm{d}w + \frac{1}{N}\,\mathrm{KL}(q(w;\theta) \parallel p(w)) = l_i(\theta) + d(\theta) \tag{5}$$

Even though this procedure makes inference analytical for a large class of models, it still lacks in many ways. This technique does not scale to large data since it requires calculations over the entire dataset and the approach does not adapt to complex models (models in which this last integral cannot be evaluated analytically) such as Bayesian neural networks [Neal, 2012, Gal, 2016]. The former challenge is tackled by [Hoffman et al., 2013] with the Stochastic Variational Inference algorithm and the latter is addressed by [Kucukelbir et al., 2017, Blundell et al., 2015] where the intractable expectation l_i in (5) is integrated by Monte Carlo. We perform this optimization step using our framework MISSO with the following quadratic surrogate at $\theta \in \Theta$:

$$f_{i,\theta}(\vartheta) \triangleq f_i(\theta) + \nabla f_i(\theta)^\top(\vartheta-\theta) + \frac{L}{2}\|\vartheta-\theta\|_2^2 \quad \text{for all } \vartheta \in \Theta \text{ and } i \in [\![N]\!] \tag{6}$$

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. Let $t : \Theta \times \mathbb{R}^d \mapsto \mathbb{R}^d$ be a function and ϕ be the density of the standard multivariate normal distribution $\mathcal{N}_d(0,\mathrm{Id})$. We assume that for all $\theta \in \Theta$, the distribution of the random vector $w = t(\theta,\epsilon)$, where $\epsilon \sim \mathcal{N}_d(0,\mathrm{Id})$, has a density $q(\cdot,\theta)$. Then, following [Blundell et al., 2015, Proposition 1], $\nabla l_i(\theta)$ is computed as:

$$\nabla l_i(\theta) = -\nabla \int_{\mathsf{W}} \log p_i(y_i|x_i,w)q(w;\theta)\mathrm{d}w = -\int_{\mathsf{W}} \mathrm{J}(\theta,z_i)\nabla \log p_i(y_i|x_i,t(\theta,z_i))\phi(z_i)\mu_i(\mathrm{d}z_i) \tag{7}$$

where for each for $i \in [\![N]\!]$, $z_i \in \mathbb{R}^d$, $\mathrm{J}(\theta,z_i)$ is the Jacobian of the function $t(\cdot,z_i)$ with respect to θ.

Thus, at iteration k we get $\theta^k = \frac{1}{N}\sum_{i=1}^N \theta^{\tau_{i,k}} - \frac{1}{2\gamma}\sum_{i=1}^N\{\hat{m}_i^k + \nabla d(\theta^{k-1})\}$ where $\hat{m}_i^k = -\frac{1}{M_k}\sum_{m=0}^{M_k-1}\mathrm{J}(\theta^{k-1},z_i^{k,m})\nabla_\theta \log p_i(y_i,x_i|t(\theta^{k-1},z_i^{k,m}))$, if $i \in I_k$, is the MC of (7) where $\{z_i^{k,m}\}_{m=0}^{M_{k-1}}$ are i.i.d samples from $\mathcal{N}_d(0,\mathrm{Id})$ and M_k is the size of the batch which might depend upon the iteration. We apply variational inference for a 2-layer Bayesian neural network on the MNIST dataset [LeCun and Cortes, 2010] using our MISSO scheme. The training set is composed of $N = 60\,000$ handwritten digits, 28×28 images, $d = 784$. Our neural network is composed of an input layer with $d = 784$ units, a single hidden layer of $p = 100$ hyperbolic tangent units and a final softmax output layer with $K = 10$ classes. We set $p(w) = \mathcal{N}(0,\mathrm{Id})$ and $p(y_i|x_i,w) = \mathrm{Softmax}(f(x_i,w))$ where f is the two layer model described above. We compare the convergence behaviors of the following state of the art optimization algorithms, using their vanilla implementations on TensorFlow [Abadi et al., 2015]: the SGD [Kiefer and Wolfowitz, 1952], the ADAM [Kingma and Ba, 2014], the SAG [Le Roux et al., 2012] and the Momentum [Sutskever et al., 2013] algorithms versus our MISSO update with a constant learning rate of 10^{-5}. Our estimator is computed using the Edward library [Tran et al., 2016]. The batch size p is set to 1% and 10% of the training set as seen in Figure 1.

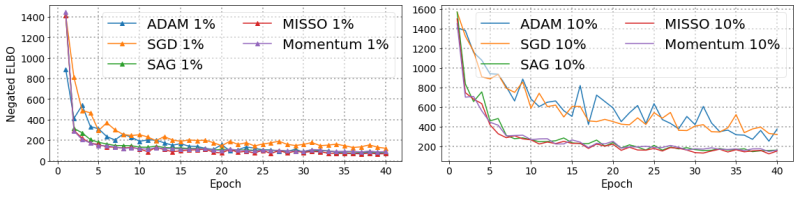

Figure 1: Convergence of the negated ELBO. Runs for two different mini-batch sizes.

Conclusion: In this paper, we have presented a unifying framework for minimization by incremental surrogate optimization when the surrogate functions are intractable and need to be approximated by MC integration. Our approach covers a large class of nonconvex optimization algorithms used in machine learning, such as mini-batch version of the Variational Inference algorithm. Non asymptotic convergence results for both convex and nonconvex objective functions can be obtained and will be reported in future works.

References

M. Abadi, A. Agarwal, P. Barham, E. Brevdo, Z. Chen, C. Citro, G.S. Corrado, A. Davis, J. Dean, M. Devin, S. Ghemawat, I. Goodfellow, A. Harp, G. Irving, M. Isard, Y. Jia, R. Jozefowicz, L. Kaiser, M. Kudlur, J. Levenberg, D. Mané, R. Monga, S. Moore, D. Murray, C. Olah, M Schuster, J Shlens, B Steiner, I. Sutskever, K. Talwar, P. Tucker, V. Vanhoucke, V. Vasudevan, F. Viégas, O. Vinyals, P. Warden, M. Wattenberg, M. Wicke, Y. Yu, and X. Zheng. TensorFlow: Large-scale machine learning on heterogeneous systems, 2015. URL `https://www.tensorflow.org/`. Software available from tensorflow.org.

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.

Z. Ghahramani. Probabilistic machine learning and artificial intelligence. *Nature*, 521, 2015.

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.

J. Kiefer and J. Wolfowitz. Stochastic estimation of the maximum of a regression function. *Ann. Math. Statist.*, 23(3):462–466, 09 1952. doi: 10.1214/aoms/1177729392. URL `https://doi.org/10.1214/aoms/1177729392`.

D.P. Kingma and J. Ba. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980, 2014. URL `http://dblp.uni-trier.de/db/journals/corr/corr1412.html#KingmaB14`.

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`.

K. Lange. MM Optimization Algorithms. 2016.

N. Le Roux, M.W. Schmidt, and F. Bach. A stochastic gradient method with an exponential convergence rate for finite training sets. In *NIPS*, pages 2672–2680, 2012.

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

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.

Konstantin Mishchenko, Franck Iutzeler, Jérôme Malick, and Massih-Reza Amini. A delay-tolerant proximal-gradient algorithm for distributed learning. In *International Conference on Machine Learning*, pages 3584–3592, 2018.

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

Y. Nesterov. Gradient methods for minimizing composite objective function. CORE Discussion Papers 2007076, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE), 2007. URL `https://EconPapers.repec.org/RePEc:cor:louvco:2007076`.

N. Parikh and S.P. Boyd. Proximal algorithms. *Foundations and Trends in Optimization*, 1(3): 127–239, 2014.

R. Ranganath, S. Gerrish, and D. Blei. Black Box Variational Inference. *PMLR*, 33:814–822, 2014.

M. Razaviyayn, M. Sanjabi, and Z. Luo. A stochastic successive minimization method for nonsmooth nonconvex optimization with applications to transceiver design in wireless communication networks. *CoRR*, abs/1307.4457, 2013.

H. Robbins and S. Monro. A stochastic approximation method. *Annals of Mathematical Statistics*, 22:400–407, 1951.

I. Sutskever, J. Martens, G. E. Dahl, and G.E. Hinton. On the importance of initialization and momentum in deep learning. In *ICML (3)*, volume 28 of *JMLR Workshop and Conference Proceedings*, pages 1139–1147. JMLR.org, 2013. URL `http://dblp.uni-trier.de/db/conf/icml/icml2013.html#SutskeverMDH13`.

D. Tran, A. Kucukelbir, A.B. Dieng, M. Rudolph, D. Liang, and D.M. Blei. Edward: A library for probabilistic modeling, inference, and criticism, 2016. URL `http://arxiv.org/abs/1610.09787`. cite arxiv:1610.09787.

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

A Proofs

Lemma 1

Let $(V_k)_{k\geq 0}$ be a non negative sequence of random variables such that $\mathbb{E}[V_0] < \infty$. Let $(X_k)_{k\geq 0}$ a non negative sequence of random variables and $(E_k)_{k\geq 0}$ be a sequence of random variables such that $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|] < \infty$. If for any $k \geq 1$:

$$V_k \leq V_{k-1} - X_k + E_k \tag{8}$$

then:

(i) for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$ and the sequence $(V_k)_{k\geq 0}$ converges a.s. to a finite limit V_∞.

(ii) the sequence $(\mathbb{E}[V_k])_{k\geq 0}$ converges and $\lim_{k\to\infty} \mathbb{E}[V_k] = \mathbb{E}[V_\infty]$.

(iii) the series $\sum_{k=0}^{\infty} X_k$ converges almost surely and $\sum_{k=0}^{\infty} \mathbb{E}[X_k] < \infty$.

Remark

Note that the result still holds if $(V_k)_{k\geq 0}$ is a sequence of random variables which is bounded from below by a deterministic quantity $M \in \mathbb{R}$.

A.1 Proof of Lemma 1

We first show that for all $k \geq 0$, $\mathbb{E}[V_k] < \infty$. Note indeed that:

$$0 \leq V_k \leq V_0 - \sum_{j=1}^{k} X_j + \sum_{j=1}^{k} E_j \leq V_0 + \sum_{j=1}^{k} E_j \tag{9}$$

showing that $\mathbb{E}[V_k] \leq \mathbb{E}[V_0] + \mathbb{E}\left[\sum_{j=1}^{k} E_j\right] < \infty$.

Since $0 \leq X_k \leq V_{k-1} - V_k + E_k$ we also obtain for all $k \geq 0$, $\mathbb{E}[X_k] < \infty$. Moreover, since $\mathbb{E}\left[\sum_{j=1}^{\infty} |E_j|\right] < \infty$, the series $\sum_{j=1}^{\infty} E_j$ converges a.s. We may therefore define:

$$W_k = V_k + \sum_{j=k+1}^{\infty} E_j \tag{10}$$

Note that $\mathbb{E}[|W_k|] \leq \mathbb{E}[V_k] + \mathbb{E}\left[\sum_{j=k+1}^{\infty} |E_j|\right] < \infty$. For all $k \geq 1$, we get:

$$\begin{aligned} W_k &\leq V_{k-1} - X_k + \sum_{j=k}^{\infty} E_j \leq W_{k-1} - X_k \leq W_{k-1} \\ \mathbb{E}[W_k] &\leq \mathbb{E}[W_{k-1}] - \mathbb{E}[X_k] \end{aligned} \tag{11}$$

Hence the sequences $(W_k)_{k\geq 0}$ and $(\mathbb{E}[W_k])_{k\geq 0}$ are non increasing. Since for all $k \geq 0$, $W_k \geq -\sum_{j=1}^{\infty} |E_j| > -\infty$ and $\mathbb{E}[W_k] \geq -\sum_{j=1}^{\infty} \mathbb{E}[|E_j|] > -\infty$, the (random) sequence $(W_k)_{k\geq 0}$ converges a.s. to a limit W_∞ and the (deterministic) sequence $(\mathbb{E}[W_k])_{k\geq 0}$ converges to a limit w_∞. Since $|W_k| \leq V_0 + \sum_{j=1}^{\infty} |E_j|$, the Fatou lemma implies that:

$$\mathbb{E}[\liminf_{k\to\infty} |W_k|] = \mathbb{E}[|W_\infty|] \leq \liminf_{k\to\infty} \mathbb{E}[|W_k|] \leq \mathbb{E}[V_0] + \sum_{j=1}^{\infty} \mathbb{E}[|E_j|] < \infty \tag{12}$$

showing that the random variable W_∞ is integrable.

In the sequel, set $U_k \triangleq W_0 - W_k$. By construction we have for all $k \geq 0$, $U_k \geq 0$, $U_k \leq U_{k+1}$ and $\mathbb{E}[U_k] \leq \mathbb{E}[|W_0|] + \mathbb{E}[|W_k|] < \infty$ and by the monotone convergence theorem, we get:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[\lim_{k\to\infty} U_k] \tag{13}$$

Finally, we have:

$$\lim_{k\to\infty} \mathbb{E}[U_k] = \mathbb{E}[W_0] - w_\infty \quad \text{and} \quad \mathbb{E}[\lim_{k\to\infty} U_k] = \mathbb{E}[W_0] - \mathbb{E}[W_\infty] \tag{14}$$

showing that $\mathbb{E}[W_\infty] = w_\infty$ and concluding the proof of (ii). Moreover, using (11) we have that $W_k \leq W_{k-1} - X_k$ which yields:

$$\begin{aligned} \sum_{j=1}^{\infty} X_j &\leq W_0 - W_\infty < \infty \\ \sum_{j=1}^{\infty} \mathbb{E}[X_j] &\leq \mathbb{E}[W_0] - w_\infty < \infty \end{aligned} \tag{15}$$

which concludes the proof of the lemma.

A.2 Proof of theorem 1

A.2.1 Proof of (i)

Set for all $\vartheta \in \Theta$, $i \in [\![N]\!]$ and $k \geq 1$:

$$a_i^k(\vartheta) \triangleq f_{i,\theta^{\tau_{i,k}}}(\vartheta) \quad \text{and} \quad \bar{a}^k(\vartheta) = \sum_{i=1}^{N} a_i^k(\vartheta) \tag{16}$$

where the function $f_{i,\theta^{\tau_{i,k}}}$ is defined by (1) and $\tau_{i,k}$ is defined recursively as follows:

$$\tau_{i,k} = k-1 \quad \text{if } i \in I_k \quad \text{and } \tau_{i,k} = \tau_{i,k-1} \text{ otherwise} \tag{17}$$

For any $k \geq 1$ and $\theta \in \Theta$ the following decomposition plays a key role:

$$\hat{a}^k(\vartheta) = \hat{a}^{k-1}(\vartheta) + \sum_{i\in I_k} \{\hat{f}_{i,\theta^{k-1}}^{M_k}(\vartheta) - \hat{a}_i^{k-1}(\vartheta)\} \tag{18}$$

where for all $\vartheta \in \Theta$, $i \in [\![N]\!]$ and $k \geq 1$:

$$\hat{a}_i^k(\vartheta) \triangleq \hat{f}_{i,\theta^{\tau_{i,k}}}^{M_k}(\vartheta) \quad \text{and} \quad \hat{a}^k(\vartheta) = \sum_{i=1}^{N} \hat{a}_i^k(\vartheta) \tag{19}$$

Set the following notations:

$$\begin{aligned} V_k &\triangleq \bar{a}^k(\theta^k), \\ X_k &\triangleq -\sum_{i\in I_k} \{f_{i,\theta^{k-1}}(\theta^{k-1}) - a_i^{k-1}(\theta^{k-1})\}, \\ E_k &\triangleq \sum_{i\in I_k} \{\hat{f}_{i,\theta^{k-1}}^{M_k}(\theta^{k-1}) - f_{i,\theta^{k-1}}(\theta^{k-1})\} \\ &\quad + \sum_{i\in I_k} \{a_i^{k-1}(\theta^{k-1}) - \hat{a}_i^{k-1}(\theta^{k-1})\} \\ &\quad + \bar{a}^k(\theta^k) - \hat{a}^k(\theta^k) + \hat{a}^{k-1}(\theta^{k-1}) - \bar{a}^{k-1}(\theta^{k-1}). \end{aligned}$$

Combining (18) with $\bar{a}^k(\theta^k) = \bar{a}^k(\theta^k) - \hat{a}^k(\theta^k) + \hat{a}^k(\theta^k)$ and $\hat{a}^k(\theta^k) \leq \hat{a}^k(\theta^{k-1})$, we obtain:

$$V_k \leq V_{k-1} - X_k + E_k. \tag{20}$$

where a_i^{k-1} and $\bar{a}^k$ are defined in (16). We now check the assumptions of Lemma 1. Note first that the sequence $(V_k)_{k\geq 0}$ is bounded from below under assumption M 1. We now check that $X_k \geq 0$ thanks to the following relation obtained using the definition of surrogate functions:

$$X_k = \sum_{i\in I_k} \{a_i^{k-1}(\theta^{k-1}) - f_{i,\theta^{k-1}}(\theta^{k-1})\} = \sum_{i\in I_k} \{a_i^{k-1}(\theta^{k-1}) - f_i(\theta^{k-1})\} \geq 0. \tag{21}$$

We finally have to prove the convergence of the series $\sum_{k=0}^{\infty} \mathbb{E}[|E_k|]$. For this purpose, we will show that for all $i \in [\![N]\!]$:

$$\sum_{k=0}^{\infty} \mathbb{E}\big[|\hat{a}_i^k(\theta^k) - a_i^k(\theta^k)|\big] < \infty \tag{22}$$

We have, using the Tower property of the conditional expectation and the Jensen inequality:

$$\mathbb{E}\big[|\hat{a}_i^k(\theta^k) - a_i^k(\theta^k)|\big] \leq \mathbb{E}\left[\mathbb{E}_{i,\theta^{\tau_{i,k}}}\left[\sup_{\vartheta\in\Theta} |\hat{a}_i^k(\vartheta) - a_i^k(\vartheta)|\right]\right] \tag{23}$$

Under assumption M 3 applied with the function $\vartheta \to \hat{a}_i^k(\vartheta)$, for all $i \in [\![N]\!]$ we have:

$$\mathbb{E}_{i,\theta^{\tau_{i,k}}}\left[\sup_{\vartheta\in\Theta} |\hat{a}_i^k(\vartheta) - a_i^k(\vartheta)|\right] \leq C_i(r_{i,\theta^{\tau_{i,k}}}) M_{\tau_{i,k}}^{-1/2} \tag{24}$$

where $C_i(r_{i,\theta^{\tau_{i,k}}})$ is a finite constant defined by (3) and $\tau_{i,k}$ is defined by (17).

Thus, we have that:

$$\mathbb{E}\big[|\hat{a}_i^k(\theta^k) - a_i^k(\theta^k)|\big] \leq C_i(r_{i,\theta^{\tau_{i,k}}}) \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] \tag{25}$$

Since, any index i is included in a mini-batch with a probability equal to $\frac{p}{N}$ conditionally independently from the past, we obtain that:

$$\mathbb{E}[M_{\tau_{i,k}}^{-1/2}] = \sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \tag{26}$$

Taking the infinite sum of this term yields:

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathbb{E}[M_{\tau_{i,k}}^{-1/2}] &= \sum_{k=1}^{\infty}\sum_{j=1}^{k} \left(1 - \frac{p}{N}\right)^{j-1} \frac{p}{N} M_{k-j}^{-1/2} \\
&= \sum_{k=1}^{\infty}\sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{k-(l+1)} \frac{p}{N} \mathbb{1}_{\{l\leq k-1\}} M_l^{-1/2} \\
&= \frac{p}{N} \sum_{l=0}^{\infty} \left(1 - \frac{p}{N}\right)^{-(l+1)} M_l^{-1/2} \sum_{k=l+1}^{\infty} \left(1 - \frac{p}{N}\right)^{k} \\
&= \sum_{l=0}^{\infty} M_l^{-1/2}
\end{aligned} \tag{27}$$

which proves identity (22), using assumption M 4. By summing over the indices $i \in [\![N]\!]$, (22) implies:

$$\sum_{k=0}^{\infty} \mathbb{E}\big[|\hat{a}^k(\theta^k) - \bar{a}^k(\theta^k)|\big] < \infty \tag{28}$$

Hence, we obtain that $\sum_{k=0}^{\infty} |\hat{a}^k(\theta^k) - \bar{a}^k(\theta^k)| < \infty$ almost surely which implies that:

$$\lim_{k\to\infty} \hat{a}^k(\theta^k) - \bar{a}^k(\theta^k) = 0 \quad \text{a.s.} \tag{29}$$

Similarly, using assumption M 3 applied for all $i \in [\![N]\!]$, with the function $\vartheta \to \nabla\hat{a}_i^k(\vartheta)$ we obtain:

$$\lim_{k\to\infty} \nabla\hat{a}^k(\theta^k) - \nabla\bar{a}^k(\theta^k) = 0 \quad \text{a.s.} \tag{30}$$

It follows from (22) and (28) that $\sum_{k=0}^{\infty} \mathbb{E}\left[|E_k|\right] < \infty$ and that the series $\sum_{k=0}^{\infty} \epsilon_k$ converges to an almost surely finite limit. Hence by Lemma 1 and (29) we get:

- the sequence $\left(\bar{a}^k(\theta^k)\right)_{k\geq 0}$ and the series $\sum_{k=0}^{\infty} \chi_k$ converge a.s.
- the sequence $\left(\mathbb{E}\left[\bar{a}^k(\theta^k)\right]\right)_{k\geq 0}$ and the series $\sum_{k=0}^{\infty} \mathbb{E}\left[X_k\right]$ converge with $\lim_{k\to\infty} \mathbb{E}\left[\bar{a}^k(\theta^k)\right] = \mathbb{E}[\lim_{k\to\infty} \bar{a}^k(\theta^k)]$.

- the sequence $\left(\hat{a}^k(\theta^k)\right)_{k\geq 0}$ converges a.s. and the sequence $\left(\mathbb{E}\left[\hat{a}^k(\theta^k)\right]\right)_{k\geq 0}$ converges.

Now, we have to prove the almost-sure convergence of the sequence $\left(f(\theta^k)\right)_{k\geq 0}$ and the convergence of $\left(\mathbb{E}\left[f(\theta^k)\right]\right)_{k\geq 0}$.

Let us denote for all $\theta \in \Theta$ and a subset $J \subset [\![N]\!]$:

$$\begin{aligned} f_J(\theta) &\triangleq \sum_{i\in J} f_i(\theta) \\ a_J^{k-1}(\theta) &\triangleq \sum_{i\in J} a_i^{k-1}(\theta) \end{aligned} \tag{31}$$

The Beppo-Levi theorem and the Tower property of the conditional expectation imply:

$$\begin{aligned} \mathsf{M} \triangleq \mathbb{E}\left[\sum_{k=1}^{\infty} X_k\right] &= \sum_{k=0}^{\infty} \mathbb{E}\left[a_{I_k}^{k-1}(\theta^{k-1}) - f_{I_k}(\theta^{k-1})\right] \\ &= \sum_{k=0}^{\infty} \mathbb{E}\left[\mathbb{E}\left[a_{I_k}^{k-1}(\theta^{k-1}) - f_{I_k}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right]\right] \end{aligned} \tag{32}$$

with $\mathbb{E}\left[f_{I_k}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right] = \frac{p}{N} f(\theta^{k-1})$ and $\mathbb{E}\left[[a_{I_k}^{k-1}(\theta^{k-1}) \,\middle|\, \mathcal{F}_{k-1}\right] = \frac{p}{N}\sum_{i=1}^N a_i^{k-1}(\theta^{k-1}) = \frac{p}{N}\bar{a}^{k-1}(\theta^{k-1})$ where $\mathcal{F}_{k-1} = \sigma(I_j, j \leq k-1)$ is the filtration generated by the sampling of the indices. We thus obtain:

$$\mathsf{M} = \frac{p}{N}\sum_{k=0}^{\infty} \mathbb{E}\left[\bar{a}^{k-1}(\theta^{k-1}) - f(\theta^{k-1})\right] = \frac{p}{N}\,\mathbb{E}\left[\sum_{k=0}^{\infty} \bar{a}^{k-1}(\theta^{k-1}) - f(\theta^{k-1})\right] < \infty \tag{33}$$

which yields to:

$$\mathbb{E}\left[\sum_{k=1}^{\infty} X_k\right] = \frac{p}{N}\,\mathbb{E}\left[\sum_{k=1}^{\infty} \{\bar{a}^{k-1}(\theta^{k-1}) - f(\theta^{k-1})\}\right] < \infty \tag{34}$$

showing that:

$$\begin{aligned} &\lim_{k\to\infty} \mathbb{E}\left[\bar{a}^k(\theta^k) - f(\theta^k)\right] = 0 \\ &\lim_{k\to\infty} \bar{a}^k(\theta^k) - f(\theta^k) = 0 \quad \text{a.s.} \end{aligned} \tag{35}$$

showing that the sequence $\left(\mathbb{E}\left[f(\theta^k)\right]\right)_{k\geq 0}$ converges and that $\left(f(\theta^k)\right)_{k\geq 0}$ converges a.s.

A.2.2 Proof of (ii)

Let us define, for all $k \geq 0$, $\bar{h}_k$ as:

$$\bar{h}^k : \vartheta \to \sum_{i=1}^N a_i^k(\vartheta) - f_i(\vartheta) \tag{36}$$

$\bar{h}^k$ is L-smooth with $L = \sum_{i=1}^N L_i$ since each of its component is L_i-smooth by definition of the surrogate functions. Using the particular parameter $\vartheta^k = \theta^k - \frac{1}{L}\nabla\bar{h}_k(\theta^k)$ we have the following classical inequality for smooth functions (cf. Lemma 1.2.3 in [Nesterov, 2007]):

$$\begin{aligned} 0 \leq \bar{h}^k(\vartheta^k) &\leq \bar{h}^k(\theta^k) - \frac{1}{2L}\|\nabla\bar{h}^k(\theta^k)\|_2^2 \\ \implies \|\nabla\bar{h}^k(\theta^k)\|_2^2 &\leq 2L\bar{h}^k(\theta^k) \end{aligned} \tag{37}$$

Using (35), we conclude that $\lim_{k\to\infty} \|\nabla\bar{h}^k(\theta^k)\|_2 = 0$ a.s. Then, the decomposition of $\langle \nabla f(\theta^k), \theta - \theta^k\rangle$ for any $\theta \in \Theta$ yields:

$$\begin{aligned} \langle \nabla f(\theta^k), \theta - \theta^k\rangle &= \langle \nabla\bar{a}^k(\theta^k), \theta - \theta^k\rangle - \langle \nabla\bar{h}^k(\theta^k), \theta - \theta^k\rangle \\ &= \langle \nabla\bar{a}^k(\theta^k) - \nabla\hat{a}^k(\theta^k), \theta - \theta^k\rangle + \langle \nabla\hat{a}^k(\theta^k), \theta - \theta^k\rangle - \langle \nabla\bar{h}^k(\theta^k), \theta - \theta^k\rangle \end{aligned} \tag{38}$$

Note that θ^k is the result of the minimization of $\hat{a}^k(\theta)$ on the constrained set Θ, therefore for all $\theta \in \Theta$, $\langle \nabla \hat{a}^k(\theta^k), \theta - \theta^k \rangle \geq 0$. Thus, we obtain, using the Cauchy-Schwarz inequality:

$$\begin{aligned} \langle \nabla f(\theta^k), \theta - \theta^k \rangle &\geq \langle \nabla \bar{a}^k(\theta^k) - \nabla \hat{a}^k(\theta^k), \theta - \theta^k \rangle - \langle \nabla \bar{h}^k(\theta^k), \theta - \theta^k \rangle \\ &\geq -\|\nabla \bar{a}^k(\theta^k) - \nabla \hat{a}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 - \|\nabla \bar{h}^k(\theta^k)\|_2 \|\theta - \theta^k\|_2 \end{aligned} \tag{39}$$

By minimizing over Θ and taking the infimum limit, we get, using (30):

$$\lim \inf_{k \to \infty} \inf_{\theta \in \Theta} \frac{\langle \nabla f(\theta^k), \theta - \theta^k \rangle}{\|\theta - \theta^k\|_2} \geq - \lim_{k \to \infty} \left(\|\nabla \bar{a}^k(\theta^k) - \nabla \hat{a}^k(\theta^k)\|_2 + \|\nabla \bar{h}^k(\theta^k)\|_2 \right) = 0 \tag{40}$$

which is the Asymptotic Stationary Point Condition (ASPC).