MISSO: MINIMIZATION BY INCREMENTAL STOCHASTIC SURROGATE OPTIMIZATION FOR LARGE SCALE NONCONVEX AND NONSMOOTH PROBLEMS

Anonymous authors
Paper under double-blind review

Reviewer 1 (3: clear rejection and 5/5 confidence):

Overall, I don't find the paper well-developed and doesn't meet the bar of a top conference like ICLR for the following major concerns: The major flaw is that in each iteration, the algorithm requires us to find the minimizer of the updated total loss (Step 8 of algorithm 2). This step is computationally as expensive as the update step in a batched MM algorithm. For a stochastic-type algorithm, I would expect the update only finds the minimizer of the stochastically picked individual surrogate function.

As the reviewer rightfully mentioned, the M-step (line 8 of Algorithm 2) is as costly as the batch MM method. The advantage of our incremental method MISSO occurs in line 7 where only a mini batch of stochastic surrogates are updated (the majority remaining unchanged). Yet, if one considers the Variational Inference example (we give page 4), the M-step is performed using stochastic VI, via stochastic gradient updates due to the quadratic nature of the surrogates. Thus, the complexity of line 8 is similar to any stochastic method (i.e., independent of n, while requiring the storage of (n-1) gradients to compute the final drift term. Likewise for the stochastic EM example, where, since the complete log likelihood belongs to the curved exponential family, the opaque M-step (line 8) is expressed with respect to the sufficient statistics, see Section B.3 of the supplementary material. Hence, the M-step actually leverages the incremental characteristic of MISSO since the stochastic sufficient statistics used in the M-step are also incrementally updated. In short, line 7 of MISSO always displays the advantage of any stochastic method in terms of computation complexity. It is also the case for the maximization step in line 8 whenever the argmin operation can be explicit (either by a gradient update or by an explicit maximization function as in the EM example).

By minimizing a stochastically picked individual surrogate function, the convergence follows by existing literature on stochastic proximal gradient method, there Theorem 2 follows without much difficulty.

First of all, we emphasize that the stochasticity in our framework (and settings in general) is double. Not only is the individual surrogate function stochastically picked (as stated by the reviewer), but the latter is also approximated by Monte Carlo sampling. This double level of stochasticity makes existing theoretical convergence proofs impractical for our work. While Theorem 2 is straightforward after establishing Theorem1, we must stress that without the latter, existing results on stochastic proximal gradient algorithms are not sufficient to prove such almost sure convergence. Indeed, we shall recall to the reviewers that our MISSO framework intends to be more general than stochastic gradient method as it also includes, among others, EM-like algorithms that can not always be casted as gradient methods. In 'Proximal-proximal-gradient method' by Ryu and Yin, the authors propose a general framework to unify proximal gradients algorithms and cast them as MM methods (for deterministic and convex objectives), yet the method uses stepsizes of order 1/L rather than n/L. Hence, we believe that the study of proximal gradient methods in our settings constitutes materials for another research paper.

The convergence rate of the proposed method is not derived, which shouldn't be too difficult to derive.

Theorem 1 is the non asymptotic convergence rate of our MISSO method. It gives a global, i.e. independent of the initialization of the algorithm, and non-asymptotic, i.e. true for any random termination number K, a rate on (1) how fast the gradient of the gap between the surrogate and the objective function decreases and (2) how fast the negative part of our stationary condition (equation

(14)) goes to zero. We explicitly write that 'the MISSO method converges to a stationary point of (1) asymptotically and at a sublinear rate $\mathbb{E}[g_{-}^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$' (see page 6). Hence, according to Theorem 1, MISSO requires $K = \mathcal{O}(nL/\epsilon)$ iterations to ensure $||g_{-}(\theta^K)|| < \epsilon$ (as a definition of ϵ-stationarity in the context of constrained optimization).

Reviewer 2 (7: good paper and 4/5 confidence):

This manuscript contributes a stochastic optimization method for finite sums where the loss function is itself an intractable expectation. It builds upon stochastic majorization-minimizations methods, in particular MISO, that it extends to use Monte-Carlo approximation of the loss. I am happy to see some attention put to the majorization-minimizations methods, which have many interesting benefits. The paper contributes nice theoretical results, in particular non-asymptotic results. However, I believe that these theoretical results are not enough to situate the contribution with regards to the wider landscape of optimization methods for machine learning. In this respect, the empirical study is crucial, however it is not completely convincing. Expressing figures 1 and 2 as a function of the number of epoch, rather than as an estimate of runtime is not meaningful: it discards the cost of running the inner loop, which varies from one approach to another. It would leed to believe that MISSO50 is the best option, which is probably not the case.

The tested methods involve similar number of gradient computations per iteration (since reported every epoch), as such the wall clock time per iteration are comparable.

Also, MC-ADAM seems to outperform MISSO for variational inference

We must acknowledge that while our MISSO scheme does not beat the SOTA (such as MC-ADAM) on every example, this paper proposes a simple yet general incremental optimization framework which encompasses several existing algorithms for large-scale data. We have tackled the challenging analysis for an algorithm with double stochasticity (index and latent variable sampling), which is not a minor contribution.

With regards to the broader contribution, it is very appreciable to have a wider theory of stochastic optimization with MM methods. It would have been good, however, to have a discussion of the link of the contributed method to the follow up work by Mairal and colleagues, Stochastic Approximate MM (Mensch et al 2017).

We believe the reviewer make a reference to 'Stochastic Subsampling for Factorizing Huge Matrices' by Mensch et. al. (https://arxiv.org/pdf/1701.05363.pdf). In this paper, the authors focus on the problem of matrix factorization for the purpose of dictionary learning. In the particular and challenging case of sparsity and high dimensional matrices (typical in fMRI data), the authors propose a stochastic MM scheme. The level of stochasticity occurs in the index sampling step (sampling subset of dimensions), see first step in Algorithm 3 in their paper, then compute the parameters of the surrogate function leveraging a deterministic (no added stochasticity in this step) Robbins-Monro type of update. Rather, in our work, two levels of stochasticity are at stake. The first one is similar as theirs, i.e. the sampling of individual indices, and the second one deals with the Monte Carlo approximation of the intractable surrogate functions (written in our illustrative examples as expectations). The theoretical and practical study of a doubly stochastic as MISSO constitute the main contributions of our paper compared to the mentioned reference, while sharing similar assumptions on the model such as smoothness and existence of directional derivative (see their assumptions (D) and (E).

Reviewer 3 (7: good paper and 3/5 confidence):

This paper propose a doubly stochastic MM method based on Monte Carlo approximation of these stochastic surrogates for solving nonconvex and nonsmooth optimization problems. The proposed method iteratively selects a batch of functions at random at each iteration and minimize the accumulated surrogate functions (which are expressed as an expectation). They establish asymptotic and non-asymptotic convergence of the proposed algorithm. They apply their method for inference of logistic regression model and for variational inference of Bayesian CNN on the real-word data sets. Weak Points. W1. The authors do not discuss the connections with state-of-the-art second-order optimization algorithms such as K-FAC. W2. The proposed algorithm still falls into the framework of MM algorithm and a simple convex quadratic surrogate function is considered. The convergence rate of the algorithm is expected. As for Reviewer 1: Theorem 1 is the non asymptotic convergence rate of our MISSO method. It gives a global, i.e. independent of the initialization of the

algorithm, and non-asymptotic, i.e. true for any random termination number K, rate. As stated page 4 'the MISSO method converges to a stationary point of (1)asymptotically and at a sublinear rate $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$'.

The research direction regarding second order surrogates is an interesting one. We can for instance think of using Newton-like updates by minimizing Hessian-informed surrogates. In 'IQN: An incremental quasi-Newton method with local superlinear convergence rate.' by Mokhtari, Eisen, and Ribeiro, a BFGS like method using memorized quantities to reduce the variance of stochastic approximations is applied to the problem of stochastic optimization leveraging quasi-Newton functions. Their work is on (a) convex and strongly convex functions and (b) deterministic surrogates. The extension to nonconvex objective and stochastic (approximated by MC) is an interesting question yet not trivial.

As for Reviewer 1: Theorem 1 is the non asymptotic convergence rate of our MISSO method. It gives a global, i.e. independent of the initialization of the algorithm, and non-asymptotic, i.e. true for any random termination number K, rate. As stated page 4 'the MISSO method converges to a stationary point of (1) asymptotically and at a sublinear rate $\mathbb{E}[g_-^{(K)}] \leq \mathcal{O}(\sqrt{1/K_{\mathsf{max}}})$'. Hence, according to Theorem 1, MISSO requires $K = \mathcal{O}(nL/\epsilon)$ iterations to ensure $||g_-(\theta^K)|| < \epsilon$ (as a definition of ϵ-stationarity in the context of constrained optimization).

Strong Points. S1. The proposed method can be viewed as a combination of MM and stochastic gradient method with variance reduction, which explains its good performance. S2. The paper contains sufficient details of the choice of the surrogate function and all the compared methods in the experiments. S3. The authors establish asymptotic and non-asymptotic convergence of the proposed algorithm. I found the technical quality is very high. S4. Extensive experiments on binary logistic regression with missing values and Bayesian CNN have been conducted.

Reviewer 4 (5 Marginally below and 1/5 confidence):

This paper proposed MISSO, which is an extension of MISO to handle surrogate functions that are expressed as an expectation. MISSO just used the Monte Carlo samples from the distribution to construct objectives to minimize. It seems to me that MISSO is just a straigforward extension of MISO, also the empirical results seems to suggest the proposed MISSO has no advantage over Monte Carlo variants of other optimizers, such as MC-SAG, MC-ADAM, thus it is not clear to me what is the significant aspect of this work.

We want to stress on the generality of our incremental optimization framework, which tackles a constrained, non-convex and non-smooth optimization problem. The main contribution of this paper is to propose and analyze a unifying framework for a large class of optimization algorithms which includes many well-known but not so well-studied algorithms. The major idea here is to relax the class of surrogate functions used in MISO [Mairal, 2015] and to allow for intractable surrogate that can only be evaluated by Monte-Carlo approximations. We provide a general algorithm and global convergence rate analysis under mild assumptions on the model and show that two examples, MLE for latent data models and Variational Inference, are its special instances. The major idea here is to relax the class of surrogate functions used in MISO [Mairal, 2015] and to allow for intractable surrogate that can only be evaluated by Monte-Carlo approximations. Working at the crossroads of Optimization and Sampling constitutes what we believe to be the novelty and the technicality of our theoretical results.

Reviewer 5 (5 Marginally below and 3/5 confidence):

(i). (Weakness) For the hard cases where each component is an expectation itself, the strategy applied here is to do a simple sample average approximation. This requires the sample size of in each iteration (M_k) to satisfy the condition that $\sum_k M_k^{-1/2} < \infty$. That is, in the k-th iteration, the sample size will be at least k^2. According to Theorem 1, the number of iteration should be $K \geq nL/\epsilon^2$. Consequently, the total sample complexity of this method seems to be $\sum_{i=1}^K k^2\, n^3 L^3 \epsilon^{-6}$. The $n^3 L^3$ dependence seems very bad. However, let us do a simple estimation of a naive method: 1. In each step compute the ϵ-accurate estimation of the gradient for each component, this needs $O(n\epsilon^{-2})$ samples per iteration. Then if the function is L-smooth (this paper can handle nonsmooth cases) then the total iterations will be O($L\epsilon^{-2}$). Then the total sample complexity seems only $O(nL\epsilon^{-4})$. This might need some clarification.

(ii). (Strength) This paper provides a non-asymptotic rate of convergence for the MISSO algorithm, which implies a non-asymptotic rate for the MISO method, whose non-asymptotic rate is not known before, which should be appreciated. Moreover, the numerical experiment in this paper is well presented. Provide additional feedback with the aim to improve the paper. Make it clear that these points are here to help, and not necessarily part of your decision assessment. (i). The MISSO (and MISO) share a similar updating style with SAG, it will be better if the authors could add some discussion on their relation and difference. Or, if such discussion exists in other literature, add a reference to that.

Indeed, the MISSO update, in the special case of quadratic surrogates, yields to a gradient update very similar to the SAG [Le Roux, Schmidt, Bach, 2012] update. Yet, the authors would like to draw the attention on the first term of the update rather than the drift term. In our method, since the minimization occurs on the aggregate sum of the stochastic surrogates, a simple derivation of the minimization of quadratic functions gives this term as being equal to the mean of the past n iterates (i.e. $1/n \sum_{i=1}^{n} \theta^{\tau_i^k}$). Whereas in SAG, as any variance reduction technique, the contribution is in the drift term (constructed through incremental update) leaving the first term unchanged vis-a-vis SGD as equal to the last iterate (i.e. θ^{k-1}). Of course when the user designed stochastic functions are no longer quadratic, the parallel with any stochastic gradient methods is no longer available making our framework more broad.

(ii). After the Theorem 2. It may make sense to give the sample complexity of the result. Namely, to get the optimality measure$\leq \epsilon$, how many sampled are needed. Specifically, by the reviewers rough estimation, the dependence on n and L is $O(n^3L^3)$, see my argument before, this dependence is not reasonable. My question is that can the authors carefully balance the parameters and derive a more reasonable sample complexity? If the $O(n)$ and $O(L)$ dependence can be achieved, the reviewer is willing to change to a higher score.

We thank the reviewer for the valuable comment on sample complexity. We give below a clarification that we hope will bring a response to the reviewer.

For starters, we first define what suboptimality condition we use in order to make a fair comparison with the proposed naive method. Let us fix that above condition as $||g_-(\theta^K)||^2$. Recall that in our case, the optimization is constrained, thus our objective function is not differentiable on the border of the constrained set Θ. Hence, the stationarity of the algorithm is characterized by the negative part of the directional derivative defined eq.(14). When $\Theta = \mathbb{R}^d$, $||g(\theta^K)||^2 = ||\nabla \mathcal{L}(\theta^K)||^2$ as typically found in unconstrained stochastic optimization literature.

Iteration Complexity: Then, according to Theorem 1, MISSO requires $K = \mathcal{O}(nL/\epsilon)$ iterations to ensure $||g_-(\theta^K)|| < \epsilon$ (ϵ-stationarity). Whereas, for the naive algorithm proposed by the reviewer, with batch setting it requires $K = L/\epsilon^2$ iterations to get ϵ-stationarity.

Sample Complexity: For the naive method, the sample complexity of nL/ϵ^4 holds. Yet, for MISSO, if we set $M_k = k^2/n^2$ such that Δ_{Kmax} is of order $\mathcal{O}(n * L)$, the sample complexity becomes $\sum_{k=0}^{nL/\epsilon} k^2/n^2 = (1/n^2) * (nL/\epsilon)^3 = nL^3/\epsilon^3$. In comparison with the proposed method (nL/ϵ^4), we sacrifice L^2 to win an order of ϵ.