Reviews of ICLR 2021

"MISSO: Minimization by Incremental Stochastic Surrogate Optimization for Large Scale Nonconvex and Nonsmooth Problems"

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

This paper develops a stochastic MM-type algorithm to minimize a finite sum. Essentially, the stochastic method draws one sample at each iteration, and find a majorization surrogate for the corresponding loss, and find the minimizer for the updated total loss.

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.

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.

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

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.

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

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).

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.

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.

Reviewer 5 (6 Marginally below and 3/5 confidence):

Summarize what the paper claims to do/contribute. Be positive and generous. In this paper, the authors consider solving the optimization of the summation of a finite number of component functions. The proposed algorithm is based on a previous work called Minimization by Incremental Surrogate Optimization (MISO). The MISO is a majorization minimization algorithm, which shares a similar update style of the SAG method. However, different from SAG, whose convergence is not available for nonconvex optimization, and is even very tricky in convex case, MISO enjoys a global convergence guarantee due to its majorization property. Based on this existing method, for the problems whose majorization surrogate is very hard to construct, e.g. variational inference of latent variable models, the authors of this paper propose a sample average approximation of the exact majorization surrogate function. The convergence of the proposed algorithm is also provided in this paper.

Clearly state your decision (accept or reject) with one or two key reasons for this choice. This paper is marginally below the acceptance threshold.

Provide supporting arguments for the reasons for the decision.

(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 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.

(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.