Rebuttal for MISSO

We sincerely thank the four reviewers for their valuable feedback. Upon acceptance, we will include in the final version (a) *improved presentation of the bounds* and (b) *an improved presentation of originality of our method.*

Reviewer 1: We thank the reviewer for the valuable comments. Regarding your remark '$||g(\theta)|| < eps$, it requires n/eps^6 samples', we kindly suggest to check page 8 our manuscript where we give a characterization of our bound in terms of iteration complexity. Indeed, our bound in Theorem 1 suggests that $\sqrt{n}/\sqrt{eps}$ samples are required to reach an eps-stationary points, which is characterized by $E||g(\theta)|| < eps$. We are not sure how you obtained an $1/eps^6$ dependency here. This $\sqrt{n}/\sqrt{eps}$ dependency is matching [Mairal, 2015] MISO bound in the case of non stochasticity, and is quite standard in several gradient-based optimization methods.

We will clarify this paragraph in the revised paper to avoid the confusion on the iteration complexity.

Reviewer 2: We thank the reviewer for the valuable feedback on our contribution. Theory-wise, the size of the Monte-Carlo batch requires a polynomial growth, i.e. $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$ as stated in Theorem 2. This assumption is required for asymptotic convergence only. In terms of non asymptotic convergence, no consideration on the batch size is done. In practice, we tune this batch size on a case by case basis. For instance, in the BNN example, a constant MC batch size equal to 10 is used.

Reviewer 3: We thank the reviewer for the thorough analysis. Our remarks are listed below:

– Comparison with non-convex optimization: In the provided cases of [Davis et al. 2018] and [Fang et al 2018], we refer the reviewer to [Mairal, 2015] for comparison between bounds as they both tackle the deterministic variant of MM scheme (indeed the surrogates are deterministic in MISO and in the two references). Regarding our case, the surrogates are no longer deterministic but rather approximated using a Monte Carlo noise. In this latter, no non asymptotic convergence bounds can be found in the literature to the best of our knowledge.

However, the paper "Non-asymptotic Analysis of Biased Stochastic Approximation Scheme" from [Karimi et. al., 2019] provides some finite time rates for a SGD scheme where the gradient is biased and stochastic. In this case, they retrieve a bound of order $log(n)/\sqrt{n}$ where n is the iteration number as they tackle the online settings. We, on the contrary, tackle the finite batch of observations setting.

– Prime notation: as we state in the notations paragraph page 3 of our paper, the prime notation defines the directional derivative of the function, see Eq(2).

Reviewer 4: We thank the reviewer for the reviews on our contribution.

– MC batch size: We admit the typo of $M_k = k^2/n^2$ given that we need $\sum_{k=0}^{\infty} M_k^{-1/2} < \infty$. It should read $M_k = \lceil k^2/n^2 \rceil$. Hence the total number of MC samples reads $O(nL^3/eps^3)$. Thank you for that.

Also, we would like to challenge the wording 'far from optimal' used by the reviewer as we believe that there are no lower bound provided by the literature under our settings. Hence, optimality of our bound can neither be affirmed or denied. Deriving a possibly matching lower bound is an interesting idea.

A comparison with deterministic method in terms of bounds can be done and we find that a plain SGD needs $O(1/eps^2)$ iteration (see [Ghadimi and Lan; 2013]; while MISSO needs $O(1/eps)$ iterations + $O(1/eps^3)$ samples. There is a trade off in our framework where the bias can be reduced faster in terms of epochs but the MC noise is larger (by construction) due to the latent sampling step.

In practice, we tune this batch size on a case by case basis. For instance, in the BNN example, a constant MC batch size equal to 10 is used and in the MLE example it is set as a polynomial of k. Reducing this growth is an option and can only be validated through experiments. Our assumptions on M_k in Theorem 2 are for the sake of the theory.

– Comparison with MCEM literature: As rightly observed by the reviewer, the MLE example in our paper serves as an illustration of a special case of our framework MISSO. By doing so, we show that using our general principle MM approach, we can retrieve the notorious MCEM algorithm (and the Variational Inference method).

In terms of convergence guarantees, the two most important references, to the best of our knowledge, are "Convergence of the Monte Carlo expectation maximization for curved exponential families" [Fort et. al., 2003] and "On Convergence Properties of the Monte Carlo EM Algorithm" [Neath., 2012]. While [Fort et. al., 2003] and [Neath., 2012] only focus on almost-sure convergence guarantees, i.e. asymptotic, we develop a finite time bounds true for any iterations, in particular the first ones where the dynamic is important. As a side note, we retrieve the rate in $M_k^{-1/2}$ as in [Fort et. al., 2003]. One important difference with [Fort et. al., 2003] is that the authors assume i.i.d. sampling for the approximation of the surrogate functions. This assumption holds only if the chains are considered to be ergodic. In our contribution we do not suppose such assumption and our bounds are true even when the latent samples have Markovian dependency. Although, we acknowledge that assumption H4 still requires to control that MC noise.

Moreover, we would like to insist on the generality of our framework, which is the main contribution of our paper. Indeed, writing the MISSO framework has the intent to make the analysis, non asymptotically or not, easier for a large class of latent data algorithms as the VI and MCEM algorithms. Retrieving bounds of such methods, if only they have been established before, is a simple validation of our framework but does not constitute the whole novelty that we propose.