We would like to thank three reviewers for their feedback. Upon acceptance, we will include in the final version (a) *clarifications on important assumptions H1-H2* and *(b) advantages of our scheme*. We first explain a few common concerns shared by **reviewer 1**, **reviewer 2**, **reviewer 3**.

●●● **Generality of MISSO**: We want to stress on the generality of our incremental framework for *non-convex* and *non-smooth* objective functions. This work aims at relaxing the class of surrogate functions used in MISO [Mairal, 2015] to accept their noisy approximations. We provide a general algorithm and global analysis under mild assumptions on the model (such as convexity of the surrogate functions and control of the Monte Carlo approximations) and show that, in particular, two problems of increasing interests in machine learning, EM and VI, are special instances.

●● **Satisfaction of assumptions**: The two assumptions H1-H2 are rather not restrictive. H1 implies that the user-designed surrogate function is above the objective and is convex (which can be constructed for most examples). H2 is based on results from empirical processes where the fluctuations between a deterministic quantity and its MC approximation can be easily bounded. Few examples and references will be included in the revised paper.

Reviewer 1: We thank the reviewer for valuable comments and references. Our point-to-point response is as follows:

Related work: The authors thank the reviewer for these related studies and will incorporate them in the revised paper. [Nitanda+, 2017] focuses on quadratic surrogate in their scheme and [Song+, 2016] considers a specific model (Boltzmann machine). Our scheme is more general in the sense that it is valid for any type of surrogate function and applies to many latent data models. The incremental update of MISSO also allows to tackle large-scale problems.

Comparison to vanilla SGD rate: The $\mathcal{O}(1/K)$ provided by the reviewer for SGD (1) is an upper bound of a different metric used in our contributions and (2) reads $\mathcal{O}(\sigma/K)$, for Empirical Risk Minimization problems, where σ is the variance of the noisy gradient, see [Ghadimi and Lan, 2013].

Advantage of MISSO: The main contribution of this paper is to propose a unifying framework for the analysis of a large class of optimisation algorithms that uses intractable surrogate functions which require to be approximated (by Monte Carlo integration for instance). While MCEM, MC-SAG, MC-ADAM are algorithms designed for different types of problem, our scheme and analysis applies to them all. Besides, there are no competitors to the MISSO scheme - at least there are no other frameworks with the same level of generality to our knowledge.

Reviewer 2: We thank the reviewer for useful comments. In addition to the summarized points above,

MISSO scheme as an extension of MISO: Regarding reviewer's comment on our contribution, we would like to emphasize on the fact that MISO and MISSO are not only for EM algorithms - those frameworks are relevant for solving non-convex, non-smooth, large-scale optimization. No comparable stochastic MM works can be found in the literature.

Additional plots and experiments: We will provide the plots of the ELBO function for the 1st experiment. For the second experiment, there are too many parameters to show since we are training a Bayesian neural network. Implementations details for each example are presented in the supplementary material. More datasets can be used for the Logistic regression (simulated datasets of binary output) and the Bayesian Neural Network (CIFAR-10).

Reviewer 3: We thank the reviewer for the comments. Please find some clarifications below:

Non-convexity and non-smoothness of the objective function: Regarding the reviewer's summary of our contribution, we would like to clarify that we are not solving a convex problem rather a non-convex and non-smooth one. From the introduction line 16-17 "the function $\mathcal{L}_i : \mathbb{R}^p \to \mathbb{R}$ is bounded from below and is (possibly) non-convex and non-smooth."

Example1 - Relation to the EM algorithm: Indeed, the Monte Carlo EM is mentioned in Figure 1 as a baseline method for training a logistic regression with missing covariates. The solution given by MISSO to the MLE in Example 1 is similar to the incremental MCEM method, where there are two levels of stochasticity. We will strengthen this connection to MCEM methods in the final version of the paper.

Remarks on the constants of Assumption H2: These are classical results from empirical processes. We did not want to spend too much time to discuss it due to space limitation. It depends on the structure of the problem such as the Lipschitzness or the i.i.d. property of the data. These bounds are typically in the order of $O(p)$ where p is the dimension of the problem. In the final version of this paper, we will include a few examples for the explicit constants.

Presentation of the bounds: The reviewer suggested using the averaged iterates $\sum_k \boldsymbol{\theta}_k / K_{\mathsf{max}}$ in the left hand side of our bounds, yet the latter optimality condition does not lead to the desirable upper bound in our case. The theorem stated is based on a random termination scheme involving the choice of an r.v. K. Random termination scheme is common in *stochastic* non-convex optimization, e.g., [Ghadimi&Lan 2013]. It has a practical advantage over the best iterate scheme which involves evaluating the gradient for $\overline{\mathcal{L}}$ - involving complex tasks such as full pass on the data and computing the objective function.