We would like to thank four reviewers for their feedback. Upon acceptance, we will include in the final version (a) *a clearer presentation of the numerical results* and (b) *missing references*. We first discuss a common concern shared by **reviewer 1**, **reviewer 2**, **reviewer 4**.

••• **Novelty of The Contribution**: We want to stress on the generality of our incremental framework, which tackles a *constrained*, *non-convex* and *non-smooth* optimization problem. The main contribution of this paper is to propose a *unifying* framework for the analysis of a large class of optimization algorithms which indeed includes well-known but not so well-studied algorithms. The major goal here is to relax the class of surrogate functions used in MISO [Mairal, 2015] and replace that by the respective Monte-Carlo approximations. We provide a general algorithm and global convergence analysis under mild assumptions on the model and show that two examples, MLE for general latent data models and Variational Inference, are its special instances.

Working at the crossroads of *Optimization* and *Sampling* constitues, we believe, the novelty and the technicality of our theoretical results.

Reviewer 1: We thank the reviewer for valuable comments and references. We would like to make the following clarification regarding the difference with MISO:

Originality: The main contribution of the paper is to extend the MISO algorithm when the surrogate fun ctions are not tractable. We motivate the need for dealing with intractable surrogate functions when nonconvex latent data models are being trained. In this case, the latent structure yields an expected surrogate functions and the nonconvexity yields an intractable expectation to compute. The only option is to build a stochastic surrogate function based on a MC approximation.

Reviewer 2: We thank the reviewer for the useful comments. Our point-to-point response is as follows:

Numerical Plots: Due to space constraints, we only presented several dimension for the logistic parameter and the mean of the latent variable. As the reviewer mentioned, we also learn the variance of those latent variables and the convergence plots of those variances will be added to the rebuttal version.

Wallclock Time:

Wallclock time per iteration is comparable for each method. Indeed the methods always only involve first order computation. Yet, we acknowledge that MISSO can present some memory bottlenecks since it requires to store n gradients through the run. This has not been a problem for the presented numerical examples

Parameter Tuning:

The baseline methods were tuned and presented to the best of their performances both with regards to their stepsize (grid search) and minibatch size. We believe your remark refers to the first numerical example (logistic regression with missing values): Regarding the stepsize, as MCEM does not have one, we indeed tuned the stepsize of SAEM. Rather than c/k, common practice is to tune a parameter α such that $\gamma_k = 1/\gamma^\alpha$. We report results for SAEM with the best α ($\alpha = 0.6$). Regarding batch size, for SAEM and MCEM both are full batch methods and the idea here is to compare different values of minibatch size for the MISSO method to see its influence on the performances.

Reviewer 3: We thank the reviewer for valuable comments and references. Please find the following precisions regarding the numerical examples:

Assumptions and Numerical Examples:

Reviewer 4: We thank the reviewer for valuable comments and the numerous related references. Our point-to-point response is as follows:

Comparison to [Murray+, 2012] and [Tran+, 2017]:

Comparison to [Kang+, 2015]:

Comparison with MC-ADAM Figure 2: