We would like to thank the 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 optimization 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 them 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 latent data models and Variational Inference, are its special instances. Working at the crossroads of *Optimization* and *Sampling* constitues what we believe to be 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 functions 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 surrogate functions can be written as an expectation due to the latent structure of the problem and the nonconvexity yields a generally intractable expectation to compute. The only option is to build a stochastic surrogate function based on a Monte Carlo 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 dimensions for the logistic parameters and the mean of the latent variable. In the final version, the variance of these latent variables and the convergence plots of those variances will be added to the supp. material. The reviewer is right that it is hard to say if the methods find the "correct" value as there are multiple local minimas for the non-convex problem of the TraumaBase experiment, in practice we found that all methods converge to the same value. We will adjust the discussions to accurately describe the findings.

Wallclock Time: As the tested methods only involve similar number of gradient computations per iteration, the wall clock time per iteration would be comparable. In the revised paper, we will provide such comparison.

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): For stepsizes, the MCEM is stepsize-less; SAEM has been hand optimized. Particularly, we have adopted the step size of $\gamma_k = 1/k^{\alpha}$ with a tuned α. We have reported results for SAEM with the best $\alpha = 0.6$. For the batch size, both SAEM and MCEM are full batch methods. We have tested different minibatch sizes for the MISSO method to examine its effect on the performances.

Reviewer 3: We thank the reviewer for valuable comments and references. We clarify the following point:

Verification of the Assumptions: For the Logistic regression with missing values, we consider that the covariance matrix is PSD and make sure in practice that its smallest eigenvalue is away from zero. For the Variational inference example, we recall the reviewer that the updates Section C.2 stem from the minimization of the quadratic functions Eq.(11) and that indeed a projection step needs to be added in order to ensure boundedness of the iterates. For illustrative purposes we did not implement the algorithm that sticks closely to our formulation but after careful consideration, the surrogate problem being a convex one on a convex closed set, the implementation will become more complex without obstructing the theory. We will provide the two variants of the algorithm in the rebuttal.

Reviewer 4: We thank the reviewer for valuable comments and references. Below we compare the 3 references you have suggested:

Comparison to [Murray+, 2012] and [Tran+, 2017]: [Murray+, 2012] is out of scope of our paper since their focus is on purely Bayesian models where the normalizing constant depends on the latent variable. In such case, since you cannot run standard MH algorithm (normalizing constants do not cancel out in the MH ratio), the authors develop a new MCMC method. [Tran+, 2017] is relevant to our paper and will be included in the rebuttal. Though, their framework is only a *full-batch* instance of our general MISSO scheme which includes incremental Variational Inference (ELBO maximization, see Example 2) but also missing values problem which is a totally different setting.

Comparison to [Kang+, 2015]: [Kang+, 2015] solely focuses on MM scheme when the surrogate functions are deterministic, *i.e.* can be computed exactly and using full batch update (versus our incremental and scalable update). Also, their analysis requires *strong convexity* of the gap between the convex surrogate and the nonconvex objective function while our analysis only requires a *smoothness* assumption, see **H2**.

Novelty of the paper: We would like to stress on the main contributions of our paper. While not beating SOTA (MC-ADAM) on a specific example, we stress that our goal is to propose a simple yet *general* incremental optimization framework which encompasses several existing algorithms for large-scale data. As a result, their theoretical analyses and implementations follow a simple and unique update rule. On the technical aspects, as stated above, the double stochasticity (index and latent variable sampling) of the optimization algorithm makes it a challenging study.