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**: 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 and analyze a **unifying framework** for a large class of optimization algorithms which includes many well-known but not so well-studied algorithms. The major idea here is to relax the class of surrogate functions used in MISO [Mairal, 2015] and to allow for intractable surrogate that can only be evaluated by Monte-Carlo approximations. We provide a general algorithm and global convergence rate 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* constitutes 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, which is not only a modest modification:
Originality: Our main contribution 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: The tested methods involve similar number of gradient computations per iteration, as such the wall clock time per iteration are comparable. In the revised paper, we will provide a comparison w.r.t. the wallclock time.
Parameter Tuning: The baseline methods were tuned and presented to the best of their performances with regard 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. We clarify the following point on the experiments:
Verification of Assumptions: Our analysis does require the parameter to be in a compact set. For the two estimation problems considered, in practice this can be enforced by restricting the parameters in a ball. In our simulation, we did not implement the algorithms that stick closely to the compactness requirement for illustrative purposes. However, we observe empirically that the parameters are always bounded. The update rules can be easily modified to respect the requirement, e.g., for logistic regression, we can use $\|\beta\| \leq R_1, \Omega \succeq \delta \mathrm{I}, \mathrm{Tr}(\Omega) \leq R_2$ and adding log-barrier functions as regularizer; for VI, we recall the surrogate functions are quadratic (Eq.(11)) and indeed a simple projection step suffices to ensure boundedness of the iterates. These variants of the algorithm will be compared in the final version.

Reviewer 4: We thank the reviewer for valuable comments. Below we address your concerns about our novelty:
Novelty w.r.t. Prior Works: Our paper differs from the 3 suggested references as we provide *an incremental optimization framework with rigorous convergence analysis*. Specifically, **[Murray+, 2012]** focuses on pure Bayesian models for which the normalizing constant *depends on the latent variable*, where the standard MH algorithm does not apply as the normalizing constants do not cancel out. An MCMC method for sampling from such distribution is proposed and is out of scope of the current paper which aims at tackling *an optimization problem*. **[Tran+, 2017]** is relevant and will be included in the final version. Though, their framework is a *full-batch* instance of our MISSO scheme which includes incremental VI (see Example 2), also the missing values problem presents a totally different challenge, in addition we provide a convergence rate analysis. **[Kang+, 2015]** focuses on full-batch MM scheme where the surrogate functions are deterministic, similar to [Razaviyayn+, 2013]. It is different from our incremental update MISSO scheme with stochastic surrogates. 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**.
Lastly, we stress that while the MISSO scheme does not beat the SOTA (such as MC-ADAM) on every example, this paper proposes a simple yet *general* incremental optimization framework which encompasses several existing algorithms for large-scale data. We have tackled the challenging analysis for an algorithm with double stochasticity (index and latent variable sampling), which is not a minor contribution.