Reviewer 1: - Result is a rather modest improvement over the MISO method of Mairal.

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 a 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:

- (Talking about providing common class of optimizers) In general, I value such contributions, but since it is not clear that this has led to any better performance of a method, the contribution is reduced.

-Why, in figure 1, do you only show the first component of delta and beta? This gives an idea of how the method performs, but not the complete picture. Also, what about Omega, since that is another parameter being learned?

- I am not so sure that the claims of *being better* (for example line 228). Overall convergence (termination) is governed by the 'worst performing' variable, so displaying a couple is not sufficient evidence. Not to mention, it is not clear to me that your method converges more quickly to the 'correct' value.

We made sure for all experiments that the estimated parameters for each method converge to the same value. This was our only reference to claim that a method is faster than the other. Then, the problem being (highly) nonconvex, indeed the estimations can get trapped in various local minima. Regardless of generalization properties of the output vector of estimated parameters, our focus through those numerical examples was to highlight faster convergence, in iteration, of our method.

- What about timings for the numerical results? Do you ever discuss the cost of the competing methods versus your approach (say, per iteration)?

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

- Your claim of better performance (with is debatable) also does not consider the fact that you have 3 different runs of your method (different batch sizes) while nothing is said about how the other methods are parameter-tuned, if they are at all. For example, one would use $gamma_k = c/k$ for SAEM with c chosen to optimize performance. Was that done?

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:

My main consideration is the models in numerical experiments seem not to satisfy the assumption in the theory. Logistic regression: Line 205-210 states that parameter beta is unconstrained and Omega is positive definite matrix.

In fact we assume boundedness in all experiments. We constraint them in l2 balls. For theta we consider PSD such that min and max eigenvalue are away from zero. In practice we did it like this.

Bayesian CNN: The update rules in Section C.2 do not consider the constraint. Does that mean we cannot say the parameters in a compact set?

In practice we implement a projection to avoid problems with variance in particular

The constraint problem on a closed convex set, it does not make a significant difference but is a bit more complex (Lagrangian) And we will provide the two versions in the appendix with the Lagrangian and the current form. We checked everything and it does not change the theory.

Reviewer 4:

Novelty of the contribution - I am not entirely sure how novel the proposed methodology is. There are several papers that deal with intractable posterior in latent variable models (Murray, I., Ghahramani, Z., & MacKay, D. (2012). MCMC for doubly-intractable distributions. arXiv preprint arXiv:1206.6848.) and variational inference (Tran, M. N., Nott, D. J., & Kohn, R. (2017). Variational Bayes with intractable likelihood. Journal of Computational and Graphical Statistics, 26(4), 873-882.).

MCMC for doubly intractable has nothing to do. It is purely bayesian and it is the case when the normalizing constant depends theta and in such case you can run standard MH so they develop new MCMC method to sample from the posterior.

It's much more general since it covers VI (where no MCMC) and missing values problem (nothing to do with elbo problems)

- I understand in this situation there is another layer of complexity arising from the surrogate function but not entirely convinced of the difficulty of modifying the existing literature on intractable posteriors.

- Moreover, there are some global convergence results (Kang, Y., Zhang, Z., & Li, W. J. (2015). On the global convergence of majorization minimization algorithms for nonconvex optimization problems. arXiv preprint arXiv:1504.07791.) on non-smooth, non-convex optimization in the context of MM. How different those results are from yours?

It's not stochastic (compute surrogate exactly) + not big data (not incremental)

- Finally, on Fig. 2(b) I see MC-ADAM outperforming MISSO. Am I missing something?