Reviewer 3jcF (6;3): A few questions: In the experiments, it is said that the batch size is . If I understand correctly, this isn't the "batch size" as we usual understand it, but rather the number of Monte Carlo samples to estimate the inner expectation, corresponding to the batch of selected functions. Am I correct? In this case, it would be helpful to call the batch size (supposing is a set of indices sampled at iteration).

In practice, is it possible to use a fixed batch size ? I'm guessing the goal of the increasing batch size is to decrease variance of the estimator along iterations. Is another way to do this possible, without requiring growing batch sizes ? (This is probably outside of the scope of the paper, but I'd be interested to know if you've thought about this.)

I'm curious about the choice of the Bayesian CNN application. I'm not familiar with such work, and am not sure it is very much used, especially on simple datasets like MNIST/CIFAR 10. It would probably more interesting/relevant to apply MISSO to training VAEs, for which the setup is similar?

Modern neural nets w/ batch norm are known to be hard to optimize with variance reduction methods. (Defazio and Bottou, 2018, https://arxiv.org/abs/1812.04529) This may explain the results for Resnet18 on CIFAR10. Under this light, it may be of interest to develop theory for the convex case (such as linear logistic regression); or at least empirically motivate the nonconvexity aspect in a significant practical example, since it's the only case considered by the theoretical analysis.

The discussion on iteration complexity is not entirely clear : first a sublinear rate for -stationarity is declared, but then using the squared stationarity metric gives a rate. It would be interesting/helpful to remind the reader that these stationarity gaps are used because in the convex case they give upper bounds on the suboptimality. Also, see the Frank-Wolfe gap developed in the Frank-Wolfe/Conditional gradient literature, which seems highly related to the gap introduced here and in (Mairal 2014). (eg Jaggi 2013, Revisiting Frank-Wolfe).

I'm willing to update my rating with clarifications/updates on the considered empirical settings. I think this would greatly benefit the paper and its diffusion.

Our reply:

Thank you very much for your detailed review. Regarding your concern about the Monte Carlo sampling scheme, as you guessed the batch size M_k is indeed the number of MC samples needed to approximate the gradient. We will distinguish it from the mini batch of indices sampled at each iteration in the revision. This number need to grow exponentially in theory but in practice we either use a linear growth in k or a simple fixed size (it is the case in the BNN example where we use the default number of 10 samples per iteration).

We decided to showcase two different examples for our MISSO scheme. Both are latent variable models, which constitutes the main settings of our contribution, and while the first one deals with a logistic regression where the latent variables are missing data, the second one is a Bayesian variant of traditional CNNs. Bayesian Deep Learning has been the focus of a distinct community working towards adapting the Bayes rule to Deep Learning. VAE is another interesting model, yet as for a BNN, VAEs are trained using Variational inference. The surrogate functions and the MISSO updates are thus identical. For the purpose

of convergence illustration, either one can be used in this case.

Regarding the convex case analysis, our method only makes sense when the surrogate functions are intractable because they are written as an intractable integral. When the structural model is convex, those expectations become tractable, hence using MISO [Mairal, 2016] is possible. Our work focuses on the models where MISO is not directly conceivable.

Reviewer UFg6 (6;4):

The method proposed in the paper appears to be new but the idea of using stochastic approximation is not new. In addition, the paper provides detailed analysis of the new method along with reasonable assumptions. I have gone through most of the proof and find no problem with the correctness. The authors also give more details on how to implement MISSO in the appendix which is appreciated. The paper is also well-written which makes it easy to follow the development of the new method.

I have the following concerns about the paper:

I do not see detailed discussion with related methods considered in the experiments. There seem to be other methods that solve the same problem in the literature so it would be best to point out the key difference between MISSO and others. Although the paper completes the picture to deal with the stochastic setting, there is not much novelty in the development of the new method as it can be seen as a combination of MISO and stochastic approximation.

Our reply:

We thank you for your consideration of our paper.

As you rightly pointed out, 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. Working at the crossroads of *Optimization*, via MISO, and *Sampling*, via Stochastic Approximation, constitutes what we believe to be the novelty and the technicality of our theoretical results.

With regard to related methods, from a practical point of view, the simplicity of MISSO's update in Algorithm 2 is enjoyable compared to various baselines such as Adam where several (possibly computationally heavy) estimations are calculated. Besides, to the best of our knowledge, the Monte Carlo variants of those baselines have not been studied and hence can not be compared theoretically to our scheme. The convergence of relatively complex algorithms such Adagrad or Adam, while adding this layer of randomness (Monte Carlo approximation), is far from easy to obtain. Hence, we added those main baselines in our experiments to at least provide an empirical comparison to MISSO.

We provide now a few related papers that are worth mentioning and that will be included in our revised paper.

[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. While the objective function is nonconvex as in our work, the construction of their surrogate functions does not imply any latent structure, inherent to the problem, and thus are easily computed and characterized for convergence purposes. 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**.

(Murray, I., Ghahramani, Z., & MacKay, D. (2012). MCMC for doubly-intractable distributions. arXiv preprint arXiv:1206.6848.)

(Tran, M. N., Nott, D. J., & Kohn, R. (2017). Variational Bayes with intractable likelihood. Journal of Computational and Graphical Statistics, 26(4), 873-882.)

(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.)

Reviewer 6f8C (4;3): I found the manuscript to be clearly written and technically sound. However, regarding the contribution, I am not sure whether this work has sufficient merit / novelty to warrant publication. The only difference between the proposed method MISSO and MISO (proposed previously) is replacing the surrogate function, which is expressed as an expectation, by the Monte Carlo integration. This modification is small and totally within the expectation. Besides, the convergence result heavily depends on the four assumptions. However, the reasonableness for the assumptions is not well discussed. For example, for H4, it is not clear to me why Cr and Cgr are finite.

Our reply:

For Cr and Cgr insist on the bracketing number.