We thank the reviewers for the valuable feedback on our submission. We will fix all the typos and latex errors.

Reviewer 1: Thank you for pointing out the shortcomings of our discussion on $\hat{k}$ in the experiments. We will improve this in the camera-ready version. **(1)** *Stochastic optimization for VI is a special case.* We agree with the reviewer that we should clarify this point when discussing related work. However, we do not view this as a fundamental weakness of the paper. Different problems require different trade-offs between, for example, precision and speed. Our evaluation metrics ($\hat{k}$, distance between the moments) reflect these specific goals. A more general study of the suite of tools we introduce applied to other problems would certainly be interesting, but is beyond the scope of our paper. **(2)** *The paper does not provide any theoretical guarantees.* See [1] for theoretical guarantees on iterate averaging. We only became aware of [1] after our submission, so we will add a substantial discussion of this paper in the camera-ready version. To avoid poor performance of iterate averaging, we check in our workflow that the variance of the iterates is finite using $\hat{k}$. The theory for $\widehat{R}$ has been thoroughly discussed in [2]. It would be interesting to use alternative $\widehat{R}$ estimators such as those in [3], which also have strong theoretical guarantees. We will be sure to clarify these points. **(3)** *The use of predicted log-likelihood on a independent test set should be also considered.* Our interest is in accurately approximating the posterior distribution. Therefore, we focus on evaluation on the quantities relevant to this goal, namely distance between moments and $\hat{k}$. However, we agree that for completeness it is useful to have predictive results, which we will add. We provide some representative results here in Table 1.

Reviewer 2: Thank you for your suggestions to improve the tables and figures. *Benefits to more recent variational inference developments?* We very much agree with this suggestion. After the submission deadline we ran additional experiments with normalizing flows which demonstrate that our $\widehat{R}$ and MCSE diagnostics indeed work well with them too. In one experiment, we used a 4-layer normalizing flow to approximate the 8-school posterior, which reduced $\hat{k}$ from 0.63 to 0.52 and the covariance error from 10.2 to 4.6 (compared to our original experiment). We will include these additional results.

		ELPD	
	Stopping Rule	LI	IA
Lin Reg	ΔELBO	-125	-162
	MCSE	-133	**-102**
Boston	ΔELBO	-90	-105
	MCSE	-81	**-79**
8-schools	ΔELBO	-6.8	-6.9
	MCSE	-6.8	**-6.7**

Table 1: Expected log predictive density (ELPD) results on held-out test data. LI = last iterate; IA = iterate average.

Reviewer 3: Thank you for your feedback. *The suggested technique seems too heuristic and lack of theoretical grounding.* Please see responses 1 and 2 to R1. We would just like to emphasize that there is a reason MCMC is so widely used: while diagnostics like $\widehat{R}$ and effective sample size are nor perfect, in practice they are quite robust, particularly when combined with checks (like the ones we use) to verify the conditions for their use (such as finite variance) hold.

Reviewer 4: Thank you for your detailed comments. **(1)** *Progress is only incremental in nature where the authors study a very specific problem of variational inference where the true posterior belonged in the variational family.* It is true that our expository example in Fig. 1 is for the case when the true posterior belongs to the variational family. Our goal with that experiment was to highlight that stochastic optimization can be unreliable in high dimensions even in this *ideal* setting. If stochastic optimisation can be problematic in such a special case, we cannot expect it to be reliable in more complicated models. Note, however, that in Sec. 3 we study many different models where the true posterior does not belong to the variational family. These results are given in Table 1 of our manuscript and confirm our findings from the idealized case. **(2)** *The proposed methods are well-known in the MCMC literature.* We agree these methods are well-known in the MCMC context. However, we do not view this as a weakness: other than iterate averaging, none of them have been used in the setting of stochastic optimization. We view adapting and validating their use in the context of stochastic optimization is a significant contribution. **(3)** *The proposed methods are heuristics.* Please see responses 1 and 2 to R1 and response 3 to R3. **(4)** *The claim that a certain stopping rule does not detect convergence is not well-supported.* We believe that Fig. 1 does clearly demonstrate this point. The inconsistency of the stopping rule is also reflected in the varying ϵ values for ΔELBO in Table 1 of our manuscript. However, we will add the experiments we did that led to these choices of ϵ in the supplementary material. **(5)** *Not enough experiments with different learning rates.* We did try varying the learning rate between 0.05 to 0.001 and did not observe a significant difference in our overall findings. We will add these results to the supplementary material. **(6)** *What distance is used is never exactly defined.* The distance is defined on line 272. We will add a forward reference when we discuss Fig. 1 in the introduction. **(7)** *$\hat{k}$ and $\widehat{R}$ are never defined.* We will add these definitions to the supplementary materials. Given limited space, we did not think the formal definitions were sufficiently enlightening to warrant inclusion in the main text.

[1] Dieuleveut, A., Durmus, A. & Bach, F. (2020). Bridging the Gap between Constant Step Size Stochastic Gradient Descent and Markov Chains. The Annals of Statistics, 48(3), 1348–1382. [2] Brooks, S., Gelman, A., Jones, G. L. & Meng, X.-L. (Eds.). (2010). Handbook of Markov Chain Monte Carlo. Chapman and Hall/CRC. [3] Vats, D. & Knudson, C. (2018). Revisiting the Gelman–Rubin Diagnostic. arXiv.org, arXiv:1812.09384 [stats.CO].