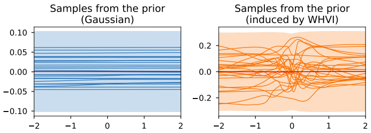
First of all we would like to thank the Reviewers for their valuable comments and useful suggestions, which will be taken into account for the revised version. With this paper we propose a novel method (Walsh-Hadamard Variational Inference) that enables VI to scale on models for which inference is known to be challenging (Bayesian DNNs and CNNs). Extensions to non-Gaussian approximate posterior (using normalizing flows) and to Gaussian processes are also highlighted in the main paper as well as in the Supplement. We are delighted to see that all the Reviewers acknowledge the novelty and the potential impact of our contribution for the NeurIPS community, as well as the overall clarity of the paper. Below, we address the main points raised by the Reviewers.

Reviewer 1: Thank you for acknowledging the significance of our contribution and its relevance for researches and pratictioners. ***"I'm not sure how well this is aligned with the goal of the paper"***. The goal is actually twofold: we show that we can reduce the parameterization of such models, and we do it in such a way that performance not only don't degradate, but they are even enhanced. For VGG16 for example, this means going from 15% of error rate with NNG down to 12% with WHVI. ***"I would have liked [...] evaluation of uncertainty calibration"***. Thanks for the suggestion and the reference. Looking at the Ovadia et al. (2019) implementation of SVI (mean-field Gaussian), we expect to behave similarly or better. ***"The paper does not compare with [...] deep ensembles [...] non-Bayesian versions of the neural networks"***. We believe deep ensembles are a bit out of the scope of the comparison but we are happy to include them in the revised paper. ***"Why did you use SGHMC instead of full HMC?"***. It was for convenience. But for this simple example we don't expect it to make big difference (the R-hat statistic showed its convergence). We will add some traces and the setup in the supplement. ***"Why do you assume a fully factorized Gaussian posterior [...]?"*** Our parameterization in case of output dimension 1 will be equivalent to mean field. WHVI can be extended to handle these cases by reshaping the parameter vector into a matrix (see experiment with GPs). ***"In which settings should one expect WHVI to work well?"*** We expect WHVI to work progressively better the deeper and the wider a model is. From the point of view of the parameterization, in the main paper and in the supplement we discuss possible limitations of the proposal and we show a numerical study where we evaluate the average distance of any random matrix to the closest WHVI matrix, showing constant behavior with increasing dimensions. ***"Table 1 is not referenced in the text"***. Thanks, there should be a reference in §2.4 ***"Caption of Figure 1 [...]. The acronym MNLL [...]. Related Work section [...]"***. Thanks for the suggestion, we will fix that.

Reviewer 2: ***"Some related work have been missed"***. Thanks for pointing out these references, which are actually very related to our work. We will include them in the final revision. ***"Setup of the CNN experiment"***. We'll point the reviewer to the supplement §D for a detailed explanation of the experimental setup. ***"The variance seems larger for WHVI [...]"***. In our experimental results, WHVI does not show systematically larger variance than other methods (see the full tabular version). For this case, we argue this is due to both having a very small dataset (the smallest, in fact) and different initializations. ***"Fig. 6 shows the calibration for WHVI?"*** Correct.

Reviewer 3: We are thrilled that you enjoyed reading our work! ***"Is not there any other better kernel approximation methods?"*** In Fig. 2 and Table 2, we have experimented with different parameterizations which can be related to different kernel approximation [58, 2]. We will make this statement more clear in the final version. ***"Why is FASTFOOD the best choice [...]?"*** As we wrote in the introduction, FASTFOOD was mostly a starting point to develop our method. Nonetheless, we tried different structures to confirm that indeed what we propose is a sensible parameterization. We believe that the superior performance comes from the sequence of projections due to the Hadamard matrix $\mathbf{H}$. Following our geometrical argument (see supplement), the Walsh-Hadamard transforms perform fast rotations of vectors living in a space of dimension D in a space $D \times D$ with complexity $D \log D$. ***"It seems that there are no answers/explanations why the proposed method can offer sensible modeling of the uncertainty"***. The cause is actually twofold: richer variational approximation and efficient parameterization. We know that from the point of view of achieving good approximations to the true posterior distribution, the mean-field family (like fully factorized Gaussian) is possibly one of the roughest approximations we can do. This is commonly done due to their simple implementation and (relative) good predictive performance. On the contrary, the posterior induced by WHVI on $\mathbf{W}$ is a matrix-variate Gaussian (in the vanilla case) with non-diagonal covariances: this is definitely more expressive than the fully factorized case.

Reviewer 4: ***"The main idea [...] is somewhat artificial from a Bayesian point of view as it would be more principled to act directly on the prior."*** Nice point! We agree that from a Bayesian p.o.v., the prior is generally the way to go. On the other hand, for these kind of models acting directly on the prior (of the parameters) is somehow very complicated and still an open problem. Recent proposals like downplaying the KL [22, 3, 48] or using temperature scaling do in fact act on the prior (implicitely, like we do) but in artificial ways (decreasing the regularization strength of the KL or increasing the magnitude of the likelihood). What we propose is developed on the solid ground of kernel methods.



As a way to characterize this behavior, we analyze the shape of the functional prior on $\mathbf{f}$. On the right, a simple visualization for a 4-layer network with 64 hidden units, where we show the pathologies of a simple Gaussian prior versus our parameterization. Thanks again for this interesting comment. We will expand on this in the final version.

"The experiments section is a little chaotic". Thanks for the suggestion, we will re-structure this section a bit, thus making it easier to follow.