| | ResNet20 | | | ResNet32 | | |
|---|---|---|---|---|---|---|
| **Model** | **ACC** | **NLL** | **ECE** | **ACC** | **NLL** | **ECE** |
| ELLA | 92.5 | 0.233 | 0.009 | 93.5 | 0.215 | 0.008 |
| Sampled LLA | 92.5 | 0.231 | 0.006 | 93.5 | 0.217 | 0.008 |
| VaLLA | 92.6 | 0.228 | 0.007 | 93.5 | 0.211 | 0.007 |
| NUQLS | 92.5 | 0.228 | 0.006 | 93.4 | 0.215 | 0.007 |

Table 1: Extension of Table 5 in our paper. Purple figures correspond to the top result, while blue figures are the second-best result. Note that our performance on ResNet20 has increased slightly due to better tuning. When tuned for predictive ability, our method can match the performance of LLA variants.

| Dataset | Method | RMSE ↓ | NLL ↓ | ECE ↓ |
|---|---|---|---|---|
| **Energy** | BDE | $0.416 \pm_{0.039}$ | $-0.125 \pm_{0.212}$ | $0.008 \pm_{0.005}$ |
| | NUQLS | $\mathbf{0.047} \pm_{\mathbf{0.006}}$ | $\mathbf{-2.400} \pm_{\mathbf{0.209}}$ | $\mathbf{0.002} \pm_{\mathbf{0.002}}$ |
| **Concrete** | BDE | $0.714 \pm_{0.054}$ | $1.563 \pm_{0.449}$ | $0.063 \pm_{0.012}$ |
| | NUQLS | $\mathbf{0.330} \pm_{\mathbf{0.047}}$ | $\mathbf{-0.316} \pm_{\mathbf{0.501}}$ | $\mathbf{0.003} \pm_{\mathbf{0.001}}$ |
| **Kin8nm** | BDE | $0.851 \pm_{0.037}$ | $1.383 \pm_{0.582}$ | $0.042 \pm_{0.012}$ |
| | NUQLS | $\mathbf{0.252} \pm_{\mathbf{0.005}}$ | $\mathbf{-0.796} \pm_{\mathbf{0.025}}$ | $\mathbf{0.000} \pm_{\mathbf{0.000}}$ |

Table 2: Comparing performance of NUQLS and BDE on UCI regression tasks. We see that NUQLS outperforms BDE on all tasks.

- *I think one of the closest methods to this paper is the Bayesian deep ensembles of He et al, though there*

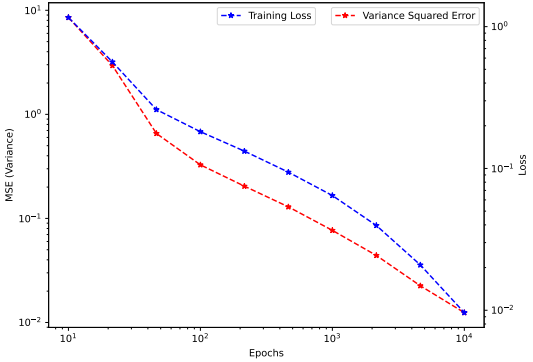
This is a very interesting idea, and we appreciate the reviewer's suggestion. We would like to explore this direction in future work. For smaller variances of $z_0$, we suspect the results of this scheme would be similar to SWAG; however, for larger initial perturbations, it might achieve performance comparable to DE. Naturally, our current theoretical analysis would no longer be valid, requiring alternative analytical approaches. Nonetheless, even in the absence of a formal theoretical foundation, we believe the empirical performance of this idea could be quite promising and worth investigating further.

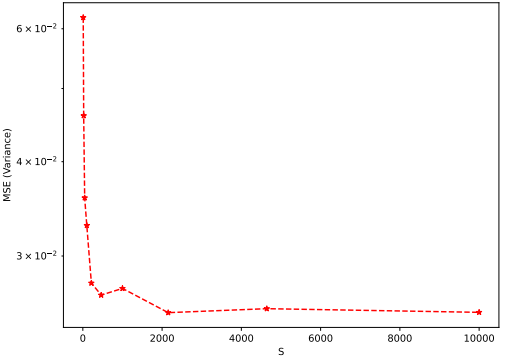
- *I don't believe the timing results for UCI datasets make sense. Specifically, I suspect for NUQLS and LLA the initial training of the NN parameters around which the approximation is centred must have not been included. DE uses 102s time, for an ensemble of 10 members, so each one must have taken 10s or so, and so I expect NUQLS and LLA will need at least 10s (and likely more, given that training a linearized model is more expensive than an unlinearized one).*

Additionally, models can be specifically engineered to excel at AUC metrics, such as the feature-space methods mentioned in our introduction, which may not necessarily reflect the true uncertainty quantification ability of the model.

We believe that uncertainty in multi-class classification should be analogous to variance in a regression setting. In regression, NLL and ECE are useful for assessing UQ ability because they are based on a heteroskedastic Gaussian distribution, which inherently includes uncertainty for each data point. We aim to replicate this by introducing a metric for uncertainty, VMSP, which is based on the variance across predictions. In future work, this variance could be leveraged to form confidence intervals, similar to Conformal intervals, to further enhance uncertainty quantification.

Finally, we note that by correctly tuning our hyperparameters, we can still achieve good predictive results (see Table 1). However, with these hyperparameters, our UQ ability, as measured by VMSP, is poor.


Training Loss
Variance Squared Error
MSE (Variance)
Loss
Epochs


MSE (Variance)
S

This paper is very interesting, and we believe it builds upon [1]. However, we argue that it is fundamentally different from our work, as the GP inference is performed with a noisy-GP on a transformed dataset. We agree that it should be included alongside the aforementioned reference in the background section. Chris: Don't discuss this, just thank them for the suggestion, say we had missed it and will cite it. They're not asking for a critique of their work...

[1] Approximate Inference Turns Deep Networks into Gaussian Processes. Khan et. al. 2019.

- *The background section of the paper is rather long and not every part is relevant. On the other hand, ablation studies on hyperparameters in supplementary are much more interesting and relevant to the key contributions of the paper. Why not move most of the section 2 to appendix, and bring back the hyperparameter analysis to experimental section?*

This is a good suggestion, as the background to our method is quite extensive and would be better

- *The metric "time" in Tables, are they run-time or training time? Is not run-time of the method more important? Also, does it also account for the time to train the MAP model? If so, how much are these?*

  We refer the reviewer to our response above to Reviewer a85w. This was unfortunately an error on our part, as we only included the inference times for the post-hoc methods. We will update the times so that they reflect both training and inference times for all methods. Since the training time is the same for NUQLS, LLA, and SWAG, and should be approximately one-tenth of the time for DE, the total time for NUQLS remains impressive in comparison to all other methods. Chris: Just post the times again if you have space, rather than redirecting to the other guy criticizing us.

# 4 Reviewer yDJq

We believe that uncertainty in multi-class classification should be analogous to variance in a regression setting. In regression, NLL and ECE are useful for assessing UQ ability as they are based on a heteroskedastic Gaussian distribution, which inherently captures uncertainty for each data point. Using VMSP, a variance term analogous to the heteroskedastic variance in Gaussian regression, we observe that NUQLS performs exceptionally well for large models and datasets, such as ResNet50 trained on CIFAR100 (see Figure 2). Moreover, by tuning our hyperparameters, we can achieve predictive performance comparable to LLA variants (see Table 1). Fred: make sure this "Table 1" is not confused with "Table 1" in the paper.

We appreciate the reviewer for raising this point, as it provides an opportunity to clarify these issues in our paper. We believe that a more in-depth discussion will enhance the paper's value, and we will incorporate it accordingly.

[1] Measuring Calibration in Deep Learning, Nixon et. al. 2020.

We are very happy to provide further clarification on this point, as the difference is indeed important. We note that the un-trainable function $\delta(.)$ in the BDE method is formed using the empirical NTK at initialization. In the infinite-width regime, this empirical NTK converges to the analytic NTK. For finite-width models however, the empirical NTK at initialization differs greatly from the empirical NTK after training (see [1]) Fred: maybe refer to issues that arise because of this, e.g., loss of feature learning, etc. We believe that the line in the BDE abstract "Finally, using finite width NNs we demonstrate that our Bayesian deep ensembles faithfully emulate the analytic posterior predictive when available..." is essentially a claim that, while the posterior analysis of BDE holds only in the infinite-width regime, their method provides a good approximation to this posterior for finite-width models. However, we find that this is not always the case in various examples (see Table 2) Fred: again, make sure this Table 2 is not confused with Table 2 in the paper.

More generally, BDE tends to incur a higher computational cost than Deep Ensembles DE due to the need to compute the untrainable function $\delta(.)$ and tune the scaling hyperparameter for classification