Addressing Bias in Active Learning with Depth Uncertainty Networks... or Not

Chelsea Murray **James U. Allingham** **Javier Antorán** **José Miguel Hernández-Lobato**
Department of Engineering
University of Cambridge
{clm88, jua23, ja666, jmh233}@cam.ac.uk

Abstract

Farquhar et al. [2021] show that correcting for active learning bias with underparameterised models leads to improved downstream performance. For overparameterised models such as NNs, however, correction leads either to decreased or unchanged performance. They suggest that this is due to an "overfitting bias" which offsets the active learning bias. We show that depth uncertainty networks operate in a low overfitting regime, much like underparameterised models. They should therefore see an increase in performance with bias correction. Surprisingly, they do not. We propose that this negative result, as well as the results Farquhar et al. [2021], can be explained via the lens of the bias-variance decomposition of generalisation error.

1 Introduction

Active learning improves data efficiency by identifying which examples maximise the expected gain in model performance [Cohn et al., 1995, Settles, 2010]. An acknowledged but relatively poorly understood problem with this approach, however, is that actively selecting informative training points introduces bias in inference, as the training data no longer follow the population distribution [MacKay, 1992]. Given that many statistical results rely on data points being identically and independently distributed (i.i.d.) samples from the population distribution, applying standard estimators to actively sampled datasets results in optimising for the wrong objective [Farquhar et al., 2021].

Recently, Farquhar et al. [2021] introduced an estimator, $\tilde{R}_{\text{LURE}}$, that corrects for the bias induced by active sampling by re-weighting the contribution of individual data points to the estimated population risk. Farquhar et al. [2021] find that, while underparameterised models such as linear regression benefit from the elimination of active learning bias, the performance of overparameterised models such as neural networks is either unaffected or negatively impacted by the unbiased risk estimator. This result is explained as active learning bias acting as a regulariser and helping to prevent flexible models from overfitting on small datasets.

Depth uncertainty networks (DUNs) are a class of Bayesian neural network (BNN) in which a prior is placed over the depth of the network, allowing us to favour simpler (i.e., shallower) functions. This enables DUNs to adapt their complexity to the varying size of the training dataset as additional labels are acquired during active learning. Because of this, we posit that DUNs will suffer less from overfitting in the early stages of training than other BNNs. This would imply that using the unbiased risk estimator of Farquhar et al. [2021] to eliminate active sampling bias should improve the performance of DUNs. Our main contributions are:

- Empirically validating the hypothesis that DUNs present less overfitting bias than other BNNs, specifically Monte Carlo dropout (MCDO), during active learning;

I (Still) Can't Believe It's Not Better Workshop at NeurIPS 2021.

- Investigating the application of the $\tilde{R}_{\text{LURE}}$ estimator to eliminate active learning bias from DUNs. Unexpectedly, despite DUNs having larger active learning bias than overfitting bias, we find that $\tilde{R}_{\text{LURE}}$ does not result in improved downstream performance;
- Providing hypotheses for why we do not observe an impact on DUNs' downstream performance when applying the unbiased risk estimator.

2 Background

2.1 Bias in active learning

We use notation similar to Farquhar et al. [2021]. Given a loss function $\mathcal{L}(\mathbf{y}, f_\theta(\mathbf{x}))$, we aim to find θ that minimises the population risk over $p_{\text{data}}(\mathbf{y}, \mathbf{x})$: $\mathbb{E}_{\mathbf{x},\mathbf{y}\sim p_{\text{data}}}\left[\mathcal{L}\left(\mathbf{y}, f_\theta(\mathbf{x})\right)\right]$. In practice, we only have access to N samples from p_{data}. These yield the empirical risk, $r = \frac{1}{N}\sum_{n=1}^{N}\mathcal{L}\left(\mathbf{y}_n, f_\theta\left(\mathbf{x}_n\right)\right)$. r is an unbiased estimator of the population risk when the data are drawn i.i.d. from p_{data}. If these samples have been used to train θ, our estimators become biased. We refer to estimators biased by computing with the training data with capital letters (i.e., R). In the active learning setting, our model is optimised using a subset of M actively sampled data points:

$$\tilde{R} = \frac{1}{M}\sum_{m=1}^{M}\mathcal{L}\left(\mathbf{y}_{i_m}, f_\theta\left(\mathbf{x}_{i_m}\right)\right), \quad i_m \sim \alpha\left(i_{1:m-1}, \mathcal{D}_{\text{pool}}\right). \tag{1}$$

The proposal distribution $\alpha\left(i_m; i_{1:m-1}, \mathcal{D}_{\text{pool}}\right)$ over the pool of unlabelled data $\mathcal{D}_{\text{pool}}$ represents the probability of each index being sampled next, given that we have already acquired $\mathcal{D}_{\text{train}} = \{\mathbf{x}_i\}_{1:m-1}$. As a result of active sampling, $\tilde{R}$ is not an unbiased estimator of R. Farquhar et al. [2021] propose $\tilde{R}_{\text{LURE}}$ ("levelled unbiased risk estimator"), which removes active learning bias:

$$\tilde{R}_{\text{LURE}} \equiv \frac{1}{M}\sum_{m=1}^{M} v_m \mathcal{L}_{i_m}; \quad v_m \equiv 1 + \frac{N-M}{N-m}\left(\frac{1}{(N-m+1)\,\alpha\left(i_m; i_{1:m-1}, \mathcal{D}_{\text{pool}}\right)} - 1\right), \tag{2}$$

where $\mathcal{L}_{i_m} \equiv \mathcal{L}\left(\mathbf{y}_{i_m}, f_\theta\left(\mathbf{x}_{i_m}\right)\right)$. Intuitively, the estimator works by re-weighting each example's contribution to the total loss by its inverse acquisition probability, such that if unusual examples have particularly high acquisition probabilities, they contribute proportionally less to the total risk. A detailed derivation of this estimator is given by Farquhar et al. [2021].

Surprisingly, Farquhar et al. [2021] find that using $\tilde{R}_{\text{LURE}}$ during training can negatively impact the performance of flexible models. They explain this by noting that the overfitting bias $(r-R)$ typical in such models has the opposite sign to the bias induced by active learning. That is, active learning encourages the selection of "difficult" to predict data points, increasing the overall risk, whereas overfitting results in the total risk being underestimated. Where overfitting is substantial, it is (partially) offset by active learning bias, such that correcting for active learning bias does not improve model performance.

2.2 Depth uncertainty networks for Active Learning

DUNs, depicted in figure 1, are a BNN variant in which the depth of the network $d \in [1, 2, ..., D]$ is treated as a random variable. Model weights θ are kept deterministic, simplifying optimisation. We place a categorical prior over depth $p(d) = \text{Cat}(d|\{\beta_i\}_{i=0}^{D})$, resulting in a categorical posterior which can be marginalised in closed form. This procedure can be seen as Bayesian model averaging over an ensemble of subnetworks of increasing depth. Thus, for DUNs, the predictive distribution and Bayesian active learning by disagreement (BALD) acquisition objective [Houlsby et al., 2011] are tractable and cheap to compute. See appendix A or [Antorán et al., 2020] for a detailed description.

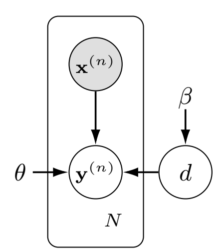

Figure 1: DUN graphical model.

In DUNs, predictions from different depths of the network correspond to different kinds of functions—shallow networks induce simple functions, while deeper networks induce more complex functions. We choose a prior that favours shallow models, limiting the flexibility of DUNs when few datapoints are observed. As the actively acquired set increases, the model likelihood will start to dominate the prior, leading the posterior over depth to favour deeper models and thus more flexible functions. We show this in Figure C.6. We hypothesize that the capability to automatically adapt model flexibility as more data are observed will reduce overfitting in the small data regimes typical of active learning. Therefore DUNs, unlike standard BNNs, may in fact benefit from the use active learning bias corrective weights. It is worth noting that automatically adapting model complexity to the observed data is also a promised feature of traditional BNNs. However, unlike in DUNs where inferences are exact, weight space models require crude approximations which fall short in this regard [Ghosh et al., 2019].

3 Results

First, using the analysis of Farquhar et al. [2021], we show that the weighted estimator removes active learning bias for DUNs. We then show that DUNs overfit less than MCDO. Finally, we examine how training with $\tilde{R}_{\text{LURE}}$ impacts downstream performance for DUNs. Experiments are performed on nine UCI regression datasets [Hernández-Lobato and Adams, 2015]; results for the Boston, Concrete and Energy datasets are presented in the following sections, with results for the rest provided in appendix C. The experimental setup is described in appendix B. All experiments are repeated 40 times, with the mean and standard deviations over the repetitions reported in our figures.

3.1 Quantifying active learning bias

We estimate the extent of active learning bias when using $\tilde{r}$ and verify the unbiasedness of $\tilde{r}_{\text{LURE}}$. A DUN is trained on $1,000$ randomly selected data points from $\mathcal{D}_{\text{pool}}$ using the standard negative log-likelihood (NLL) loss. We then estimate the model's risk using both $\tilde{r}$ and $\tilde{r}_{\text{LURE}}$, drawing M samples from an unobserved test set $\mathcal{D}_{\text{test}}$. The active learning bias is estimated as

$$B_{\text{ALB}}(\bullet) = r - E_{x_{1:M} \sim \alpha(x_{1:M};\mathcal{D}_{\text{pool}})}[\bullet], \tag{3}$$

where $\bullet$ is either $\tilde{r}$ or $\tilde{r}_{\text{LURE}}$ and r is computed on the whole $\mathcal{D}_{\text{test}}$. Figure 2 shows that $\tilde{r}_{\text{LURE}}$ is unbiased, whereas the bias in $\tilde{r}$ approaches zero only as $M \to N$.

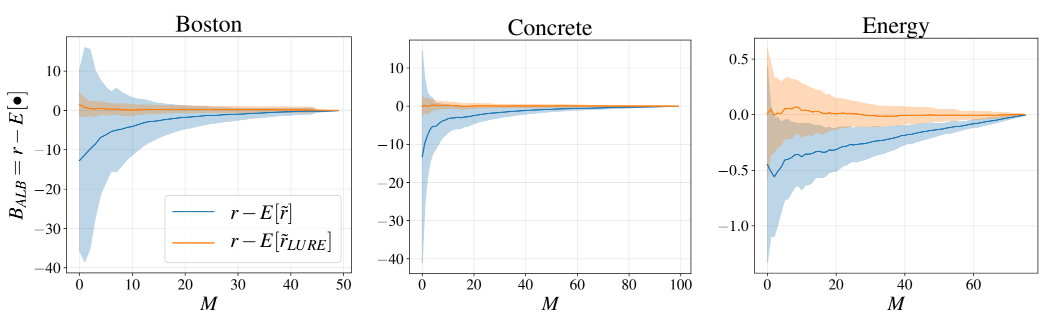

Figure 2: Active learning bias in risk estimation with DUNs, evaluated using $\tilde{r}$ (blue) and $\tilde{r}_{\text{LURE}}$ (orange). The unbiased risk estimator $\tilde{r}_{\text{LURE}}$ eliminates active learning bias, while the standard estimator $\tilde{r}$ tends to overestimate risk due to α prioritising more difficult to fit points.

3.2 Overfitting bias

To test the hypothesis that DUNs' prior over depth reduces overfitting in low data regimes, we follow the procedure of Farquhar et al. [2021] to isolate overfitting bias. We train a model with M actively sampled points using $\tilde{R}_{\text{LURE}}$ as the training objective. After acquiring each new point, we evaluate

$$B_{\text{OFB}} = r - \tilde{R}_{\text{LURE}}. \tag{4}$$

The difference between the risk evaluated on the test set (r) and on the train set once the active learning bias is removed ($\tilde{R}_{\text{LURE}}$) is taken as a measure of overfitting (B_{OFB}).

Figure 3 shows that the magnitude of B_{OFB} for DUNs is small relative to BNNs trained with MCDO [Gal and Ghahramani, 2016], one of the methods used in Farquhar et al. [2021]. This suggests that DUNs are, indeed, robust to overfitting on small datasets. Overfitting bias in DUNs is also larger in magnitude than active learning bias for the datasets shown (comparing figure 3 to figure 2). This is not the case for MCDO (comparing figure 3 to figure C.2).

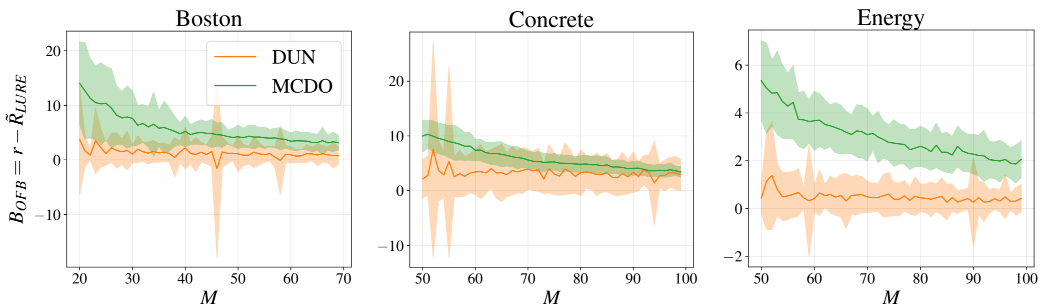

Figure 3: Evolution of overfitting bias for DUNs (orange) and MCDO (green) trained with $\tilde{R}_{\text{LURE}}$ as the amount of acquired data M grows. Overfitting bias is smaller in DUNs than for MCDO.

3.3 Training with unbiased risk estimators

According to Farquhar et al. [2021], removing the active learning bias in cases where $|B_{\text{ALB}}| \gg |B_{\text{OFB}}|$ should positively affect performance. Indeed, they observe this for linear models. Thus, having confirmed that the LURE weights eliminate active learning bias in DUNs, and that DUNs are in the $|B_{\text{ALB}}| \gg |B_{\text{OFB}}|$ regime, at least for certain datasets, we evaluate whether using the unbiased estimator during training improves downstream performance. We compare the NLL obtained by evaluating our models on $\mathcal{D}_{\text{test}}$ when using $\tilde{R}$ and $\tilde{R}_{\text{LURE}}$ as the training objectives. Surprisingly, figure 4 shows that using $\tilde{R}_{\text{LURE}}$ does not affect NLL performance at all.

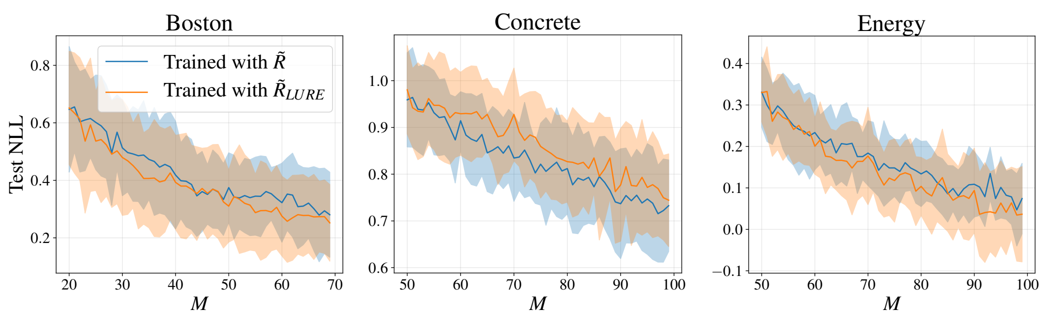

Figure 4: Test NLL as more training points are acquired for DUNs trained with $\tilde{R}$ (blue) and $\tilde{R}_{\text{LURE}}$ (orange). Active learning bias correction has minimal impact on downstream performance.

4 Discussion

In figure 4, we measure our models' generalisation error, given by $\texttt{bias}^2 + \texttt{variance}$ [Bishop, 2007]. Considering that neural networks are very flexible, we expect model bias to be negligible, and active learning bias to dominate. When we eliminate the latter with importance weights, as in equation (2), we pay the price of additional variance [Owen, 2013]. We therefore hypothesise that our results can be explained by gains due to bias reduction being lost due to increased variance. This increase in variance will depend on the choice of sampling distribution α. "Overfitting bias", as presented in equation (4) and [Farquhar et al., 2021] is computed using the train set and thus is not directly comparable with active learning bias in equation (3). Furthermore, the effect of "overfitting bias" is captured by model variance. The error induced by model variance is always additive with any statistical biases that may exist, precluding any cancellation effects. This hypothesis also explains why Farquhar et al. [2021] find $\tilde{R}_{\text{LURE}}$ to be successful for linear models: since inflexible models are well-specified by small amounts of data, applying $\tilde{R}_{\text{LURE}}$ barely increases variance. We also highlight that in the experimental design used by Farquhar et al. [2021] to obtain their linear model results, the temperature of the proposal distribution ($T = 10,000$) is much larger than that used here ($T = 10$). Higher temperatures approach a deterministic proposal, interfering with the assumption that $\alpha(\cdot)$ has full support.

References

Javier Antorán, James Allingham, and José Miguel Hernández-Lobato. Depth uncertainty in neural networks. *Advances in Neural Information Processing Systems*, 33, 2020.

Christopher M. Bishop. *Pattern Recognition and Machine Learning, 5th Edition*. Information science and statistics. Springer, 2007.

David A Cohn, Zoubin Ghahramani, and Michael I Jordan. Active learning with statistical models. Technical report, MASSACHUSETTS INST OF TECH CAMBRIDGE ARTIFICIAL INTELLIGENCE LAB, 1995.

A D'Amour, K Heller, D Moldovan, B Adlam, B Alipanahi, A Beutel, C Chen, J Deaton, J Eisenstein, MD Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. arxiv 2020. *arXiv preprint arXiv:2011.03395*, 2020.

Sebastian Farquhar, Yarin Gal, and Tom Rainforth. On statistical bias in active learning: How and when to fix it. *International Conference on Learning Representations*, 2021.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, pages 1050–1059. PMLR, 2016.

Soumya Ghosh, Jiayu Yao, and Finale Doshi-Velez. Model selection in bayesian neural networks via horseshoe priors. *Journal of Machine Learning Research*, 20(182):1–46, 2019. URL `http://jmlr.org/papers/v20/19-236.html`.

José Miguel Hernández-Lobato and Ryan Adams. Probabilistic backpropagation for scalable learning of Bayesian neural networks. In *International Conference on Machine Learning*, pages 1861–1869. PMLR, 2015.

Neil Houlsby, Ferenc Huszar, Zoubin Ghahramani, and Máté Lengyel. Bayesian active learning for classification and preference learning. *CoRR*, abs/1112.5745, 2011. URL `http://arxiv.org/abs/1112.5745`.

S. Ioffe and C. Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

David JC MacKay. Information-based objective functions for active data selection. *Neural computation*, 4(4):590–604, 1992.

Art B. Owen. *Monte Carlo theory, methods and examples*. 2013.

Burr Settles. Active learning literature survey. *Machine Learning*, 15(2):201–221, 2010.

Matthew D. Zeiler and Rob Fergus. Visualizing and understanding convolutional networks. In David J. Fleet, Tomás Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision - ECCV 2014 - 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part I*, volume 8689 of *Lecture Notes in Computer Science*, pages 818–833. Springer, 2014.