Weighted Confidence Ensemble for Uncertainty Quantification in Physics-Informed Neural Networks

Anshuman Pasupalak [1 2]
Ludovic Chamoin [3] **Massimo Pica Ciamarra** [1 4]
[1]*Division of Physics and Applied Physics, School of Physical and Mathematical Sciences, Nanyang Technological University, 21 Nanyang Link, Singapore 637371, Singapore* [2]*CNRS@CREATE, 1 Create Way, 08-01 Create Tower, Singapore 138602* [3]*ENS Paris-Saclay, LMPS, 4 Avenue des Sciences, 91190 Gif sur Yvette, France* [4]*CNR-SPIN, Dipartimento di Scienze Fisiche, Università di Napoli Federico II, I-80126 Napoli, Italy.*. Correspondence to: Massimo Pica Ciamarra massimo@ntu.edu.sg, Ludovic Chamoin ludovic.chamoin@ens-paris-saclay.fr.

1. Introduction

Machine learning (ML) has achieved remarkable performance in tasks such as image recognition [1] and natural language processing [2], fueling its widespread commercial deployment [3]. However, in critical domains like healthcare [4, 5], engineering [6, 7], and autonomous systems, the inherent 'black-box' nature of ML models poses challenges for estimating prediction uncertainty—a key factor for robust decision-making [8, 9, 10].

Uncertainty in ML is broadly categorized into *aleatoric* uncertainty, which stems from data noise, and *epistemic* uncertainty, arising from insufficient training data in or out of the domain [11]. Various strategies have been developed to quantify these uncertainties. Bayesian models [12] and Mean Variance Estimation methods [13, 14] provide principled approaches but often suffer from high computational cost [11, 15] or struggle with data sparsity [16, 17, 18]. Ensemble-based methods [19, 20] offer a practical alternative by averaging predictions from multiple models; however, they may misestimate epistemic uncertainty if individual models lack exposure to diverse data [21, 20]. Moreover, many calibration techniques cannot be applied post-training [22, 23, 24].

In this work, we propose the *Weighted Confidence Ensemble* (WCE) model to estimate both aleatoric and epistemic uncertainties in a post-training setting. Our approach is based on the observation that a network's accuracy on a test example is related to its similarity to the training data [25, 26, 27]. By applying principal component analysis (PCA) to a selected layer's activations, we quantify this similarity and derive a confidence score that reflects the network's certainty. Leveraging an ensemble of models, each providing a confidence score for its predictions, we compute uncertainty using a weighted standard deviation. This framework enables effective quantification of aleatoric uncertainty while identifying and rejecting predictions associated with high epistemic uncertainty or out of domain data.

2. Methods

2.1 Dataset

We demonstrate the efficacy of the *Weighted Confidence Ensemble* (WCE) method on both classification and regression tasks. For classification, we derive several datasets from MNIST [28] of handwritten digits. For regression, we generate a synthetic dataset based on an underdamped pendulum, where the model recovers frequency and damping constants.

#	Digits	Noise	Uncertainty
B_1, B_2	even	$p_{\text{switch}} = 0$	None
A_1	even	$p_{\text{switch}} = 0.05$	Aleatoric
A_2	even	$p_{\text{switch}} = 0.1$	Aleatoric
A_3	even	$p_{\text{switch}} = 0.2$	Aleatoric
E_1	odd	$p_{\text{switch}} = 0$	Epistemic
E_2	even	$p_{\text{switch}} = 1$	Epistemic

Table 1: MNIST-derived test datasets. The training set B_1 comprises even digits without noise.

2.1.1 Classification Datasets

The training dataset, B_1, comprises only even-digit images from MNIST. For testing, we construct multiple datasets as summarized in Table 1. Dataset B_2 contains even digits with no noise and serves as a benchmark. To simulate aleatoric uncertainty, we generate datasets A_1, A_2, and A_3 by randomly switching (black $\leftrightarrow$ white) pixel values in B_2 with probabilities $p_{\text{switch}} = 0.05, 0.1$, and 0.2, respectively. Epistemic uncertainty is introduced with dataset E_1, which contains odd digits, and dataset E_2, derived from B_2 by complete color inversion ($p_{\text{switch}} = 1$).

2.1.2 Regression Datasets

For the regression task, we consider an underdamped pendulum governed by:

$$\ddot{\theta} = -\omega^2 \sin\theta - 2\gamma\dot{\theta}, \tag{1}$$

where ω is the natural frequency and γ is the damping constant. We generate dataset D_0 by numerically integrating Eq. (1) for $t \in (0, 20)$ with $dt = 0.05$. The parameters are uniformly sampled: $\gamma \in (0.01, 0.11)$ and $\omega \in (1, 1.5)$. Each element records $\theta(t)$, $\dot{\theta}(t)$, and $\ddot{\theta}(t)$ as features, with ω and γ as labels.

The training dataset D_1 is obtained by perturbing the features and labels of D_0 with Gaussian noise. For instance, $\theta(t)$ is perturbed with a noise level of $0.1\sigma_{\theta(0)}$, where $\sigma_{\theta(0)}$ is the standard deviation of the initial condition $\theta(0)$ over D_0.

2.2 *Epistemic Uncertainty Confidence Score and Ensemble*

To quantify the epistemic uncertainty of a test input, we compare its representation on a chosen network layer with that of the training data using Principal Component Analysis (PCA) [29]. The underlying hypothesis is that a test input similar to training data will require a similar number of PCA components to capture 95% of its variance, whereas inputs deviating from the training distribution will require more, or fewer components. Let $n(x)$ denote the number of PCA components required for input x. We define the Epistemic Uncertainty Confidence Score as:

$$\mathcal{C}_{\text{ep}}(x_t) = 1 - \left(\frac{P_{50}(n(x_j)) - n(x_t)}{P_{10}(n(x_j)) - P_{90}(n(x_j))} \right)^2, \quad x_j \in T, \tag{2}$$

where $P_a(n(x_j))$ denotes the a^{th} percentile of the number of components computed on the training set T. Values of $\mathcal{C}_{\text{ep}}(x_t) < 1$ indicate that the test input x_t deviates from the training data, implying higher epistemic uncertainty.

We integrate this uncertainty measure into an ensemble framework. For each test input x, each model m in an ensemble of M neural networks produces a prediction y_p^m and an epistemic uncertainty score $\mathcal{C}_{\text{ep}}^m(x)$ as defined above. For regression tasks, the ensemble prediction is computed as:

$$y_t = \frac{\sum_{m=1}^{M} \mathcal{C}_{\text{ep}}^m(x)\, y_p^m}{\sum_{m=1}^{M} \mathcal{C}_{\text{ep}}^m(x)},$$

thus assigning greater weight to predictions with lower epistemic uncertainty. For classification tasks, we take the modal prediction and compute an overall ensemble confidence score as:

$$\mathcal{C}_{\text{ep}}^{\text{en}} = \langle \mathcal{C}_{\text{ep}}^m(x) \rangle.$$

A low $\mathcal{C}_{\text{ep}}^{\text{en}}$ indicates that the test input is out-of-distribution, exhibiting high epistemic uncertainty.

3. Results

Figure 1(a) shows the distribution of the ensemble epistemic confidence score, $\mathcal{C}_{\text{ep}}^{\text{en}}$, for the test datasets listed in Table 1. The probability distribution $P(\mathcal{C}_{\text{ep}}^{\text{en}})$ peaks at 1 for datasets similar to the training data (B_1), and the peak diminishes as the test data deviate from B_1. In particular, datasets E_1 and E_2, which significantly deviate from the training set, exhibit predominantly low $\mathcal{C}_{\text{ep}}^{\text{en}}$ values (near 0), demonstrating that the confidence score effectively captures deviations from the training data.

Figure 1(b) illustrates that the Expected Calibration Error (ECE) drops to zero for $\mathcal{C}_{\text{ep}}^{\text{en}} \simeq 1$ for all datasets except E_1 and E_2. Since most examples in E_1 and E_2 yield $\mathcal{C}_{\text{ep}}^{\text{en}} \ll 1$, the Weighted Confidence Ensemble (WCE) can reliably reject them as outliers.

Similarly, Fig. 2 shows that for the regression problem, in-domain examples peak at $\mathcal{C}_{\text{ep}}^{\text{en}} \sim 1$, while out-of-domain examples peak at approximately 0.7.

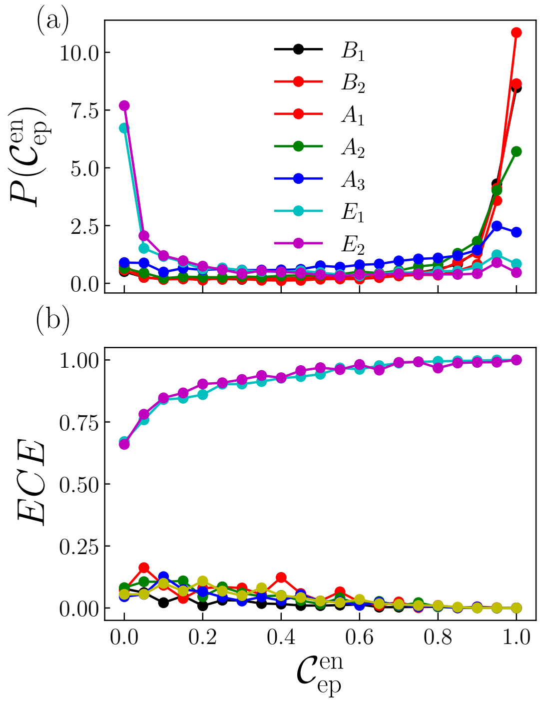

Fig. 1: Expected Calibration Error and distribution of the ensemble epistemic confidence score $\mathcal{C}_{\text{ep}}^{\text{en}}$.

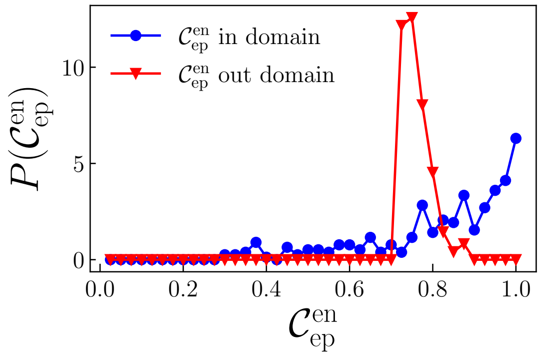

Fig. 2: Distribution of $\mathcal{C}_{\text{ep}}^{\text{en}}$ for in-domain and out-of-domain examples in the regression task.

By discarding predictions with $\mathcal{C}_{\text{ep}}^{\text{en}} < 0.85$, WCE effectively eliminates unreliable predictions.

4. Conclusion

Overall, the Weighted Confidence Ensemble (WCE) provides a robust and scalable framework for uncertainty quantification that is applicable post-training. By integrating PCA-derived confidence metrics into the ensemble prediction process, WCE effectively distinguishes between aleatoric noise and epistemic uncertainty. This dual capability enables accurate calibration of uncertainty bounds and the selective rejection of unreliable predictions, thereby enhancing the safety and interpretability of machine learning models in critical domains.

Acknowledgments

DesCartes: this research is supported by the National Research Foundation, Prime Minister's Office, Singapore under its Campus for Research Excellence and Technological Enterprise (CREATE) programme.

References

[1] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778, 2016.

[2] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[3] Nestor Maslej, Loredana Fattorini, Ray Perrault, Vanessa Parli, Anka Reuel, Erik Brynjolfsson, John Etchemendy, Katrina Ligett, Terah Lyons, James Manyika, Juan Carlos Niebles, Yoav Shoham, Russell Wald, and Jack Clark. Artificial intelligence index report 2024. *ArXiv*, abs/2405.19522, 2024.

[4] Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, David Sculley, Sebastian Nowozin, Joshua Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. *Advances in neural information processing systems*, 32, 2019.

[5] Eric J Topol. High-performance medicine: the convergence of human and artificial intelligence. *Nature medicine*, 25(1):44–56, 2019.

[6] Olufemi A Omitaomu and Haoran Niu. Artificial intelligence techniques in smart grid: A survey. *Smart Cities*, 4(2):548–568, 2021.

[7] Tanveer Ahmad, Dongdong Zhang, Chao Huang, Hongcai Zhang, Ningyi Dai, Yonghua Song, and Huanxin Chen. Artificial intelligence in sustainable energy industry: Status quo, challenges and opportunities. *Journal of Cleaner Production*, 289:125834, 2021.

[8] Wojciech Samek and Klaus-Robert Müller. Towards explainable artificial intelligence. *Explainable AI: interpreting, explaining and visualizing deep learning*, pages 5–22, 2019.

[9] Cynthia Rudin. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature machine intelligence*, 1(5):206–215, 2019.

[10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[11] Christopher KI Williams and Carl Edward Rasmussen. *Gaussian processes for machine learning*, volume 2. MIT press Cambridge, MA, 2006.

[12] Roman Garnett. *Bayesian optimization*. Cambridge University Press, 2023.

[13] David A Nix and Andreas S Weigend. Estimating the mean and variance of the target probability distribution. In *Proceedings of 1994 ieee international conference on neural networks (ICNN'94)*, volume 1, pages 55–60. IEEE, 1994.

[14] Laurens Sluijterman, Eric Cator, and Tom Heskes. Optimal training of mean variance estimation neural networks. *Neurocomputing*, page 127929, 2024.

[15] Andrew Gordon Wilson, Christoph Dann, and Hannes Nickisch. Thoughts on massively scalable gaussian processes. *arXiv preprint arXiv:1511.01870*, 2015.

[16] Georgios Papadopoulos, Peter J Edwards, and Alan F Murray. Confidence estimation methods for neural networks: A practical comparison. *IEEE transactions on neural networks*, 12(6):1278–1287, 2001.

[17] Richard Dybowski and Stephen J Roberts. Confidence intervals and prediction intervals for feed-forward neural networks. Cambridge University Press, 2001.

[18] Abbas Khosravi, Saeid Nahavandi, Doug Creighton, and Amir F Atiya. Comprehensive review of neural network-based prediction intervals and new advances. *IEEE Transactions on neural networks*, 22(9):1341–1356, 2011.

[19] Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30, 2017.

[20] HM Dipu Kabir, Abbas Khosravi, Mohammad Anwar Hosen, and Saeid Nahavandi. Neural network-based uncertainty quantification: A survey of methodologies and applications. *IEEE access*, 6:36218–36234, 2018.

[21] Stanislav Fort, Huiyi Hu, and Balaji Lakshminarayanan. Deep ensembles: A loss landscape perspective. *arXiv preprint arXiv:1912.02757*, 2019.

[22] Apostolos F Psaros, Xuhui Meng, Zongren Zou, Ling Guo, and George Em Karniadakis. Uncertainty quantification in scientific machine learning: Methods, metrics, and comparisons. *Journal of Computational Physics*, 477:111902, 2023.

[23] Hanjing Wang and Qiang Ji. Epistemic uncertainty quantification for pre-trained neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11052–11061, 2024.

[24] Achim Hekler, Titus J Brinker, and Florian Buettner. Test time augmentation meets post-hoc calibration: uncertainty quantification under

real-world conditions. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 37, pages 14856–14864, 2023.

[25] Dávid Bajusz, Anita Rácz, and Károly Héberger. Why is tanimoto index an appropriate choice for fingerprint-based similarity calculations? *Journal of cheminformatics*, 7:1–13, 2015.

[26] Marko Toplak, Rok Mocnik, Matija Polajnar, Zoran Bosnic, Lars Carlsson, Catrin Hasselgren, Janez Demsar, Scott Boyer, Blaz Zupan, and Jonna Stalring. Assessment of machine learning reliability methods for quantifying the applicability domain of qsar regression models. *Journal of chemical information and modeling*, 54(2):431–441, 2014.

[27] Lior Hirschfeld, Kyle Swanson, Kevin Yang, Regina Barzilay, and Connor W Coley. Uncertainty quantification using neural networks for molecular property prediction. *Journal of Chemical Information and Modeling*, 60(8):3770–3780, 2020.

[28] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[29] Hervé Abdi and Lynne J Williams. Principal component analysis. *Wiley interdisciplinary reviews: computational statistics*, 2(4):433–459, 2010.