How disentangled are your classification uncertainties?

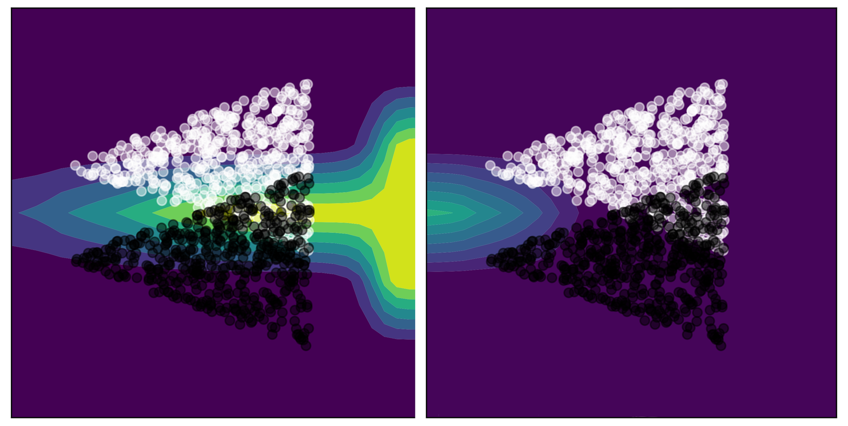
Ivo Pascal de Jong, Andreea Ioana Sburlea, Matias Valdenegro-Toro
Department of Artificial Intelligence, Bernoulli Institute, University of Groningen
ivo.de.jong@rug.nl

Abstract: Uncertainty can be separated into aleatoric (data) and epistemic (model) uncertainty. We compare *two methods* and test whether epistemic (EU) is independent of aleatoric (AU) uncertainty in *three experiments*.
We find that disentanglement does not work.

Method 1. Information Theoretic (IT) Disentanglement

$$\underbrace{\mathbb{I}(Y;\Theta|x,D)}_{Epistemic} \approx \underbrace{\mathbb{H}[\mathbb{E}_{\Theta}[p(y|x,\theta)]]}_{Total} - \underbrace{\mathbb{E}_{\Theta}[\mathbb{H}[p(y|x,\theta)]]}_{Aleatoric} \quad (1)$$

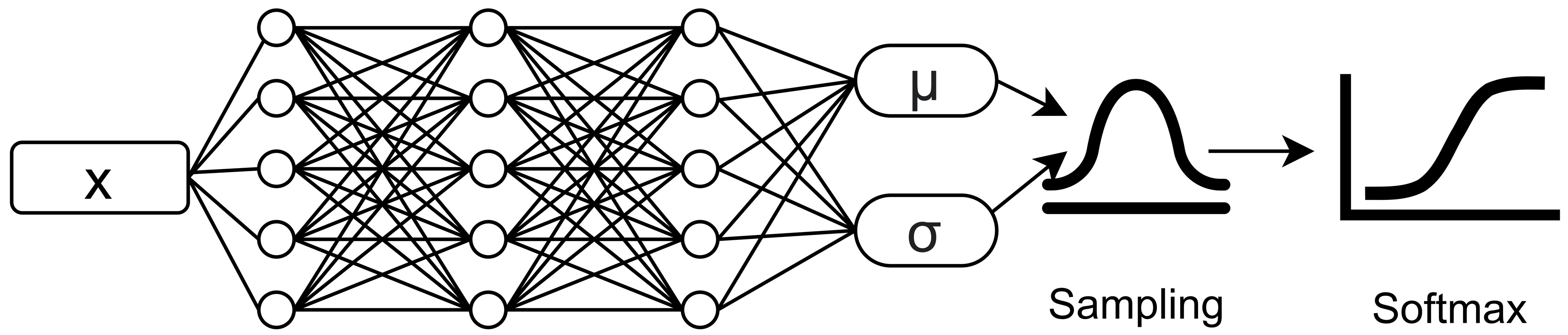
IT disentanglement [1] is the commonly used method for disentanglement. It is easy to implement, but EU is computed with a **very rough approximation**, which could be a problem in practice.



(a) (b)

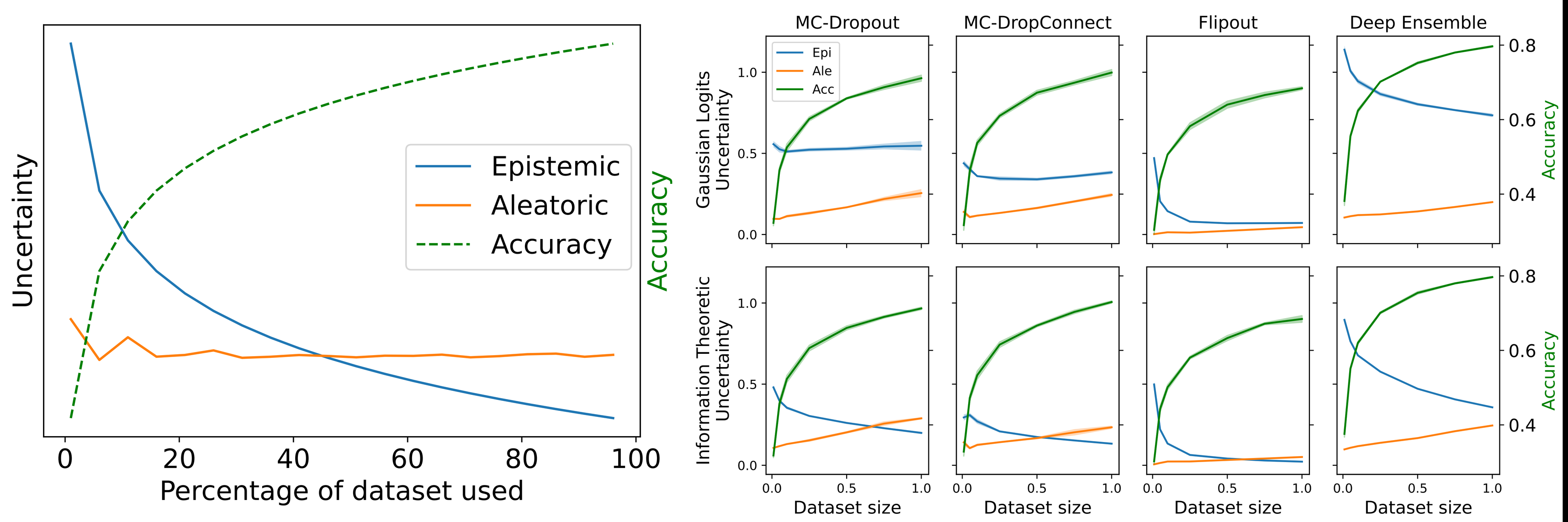
AU (a) and EU (b) using IT. With high AU, EU is underestimated. [2]

Method 2. Gaussian Logits (GL) Disentanglement



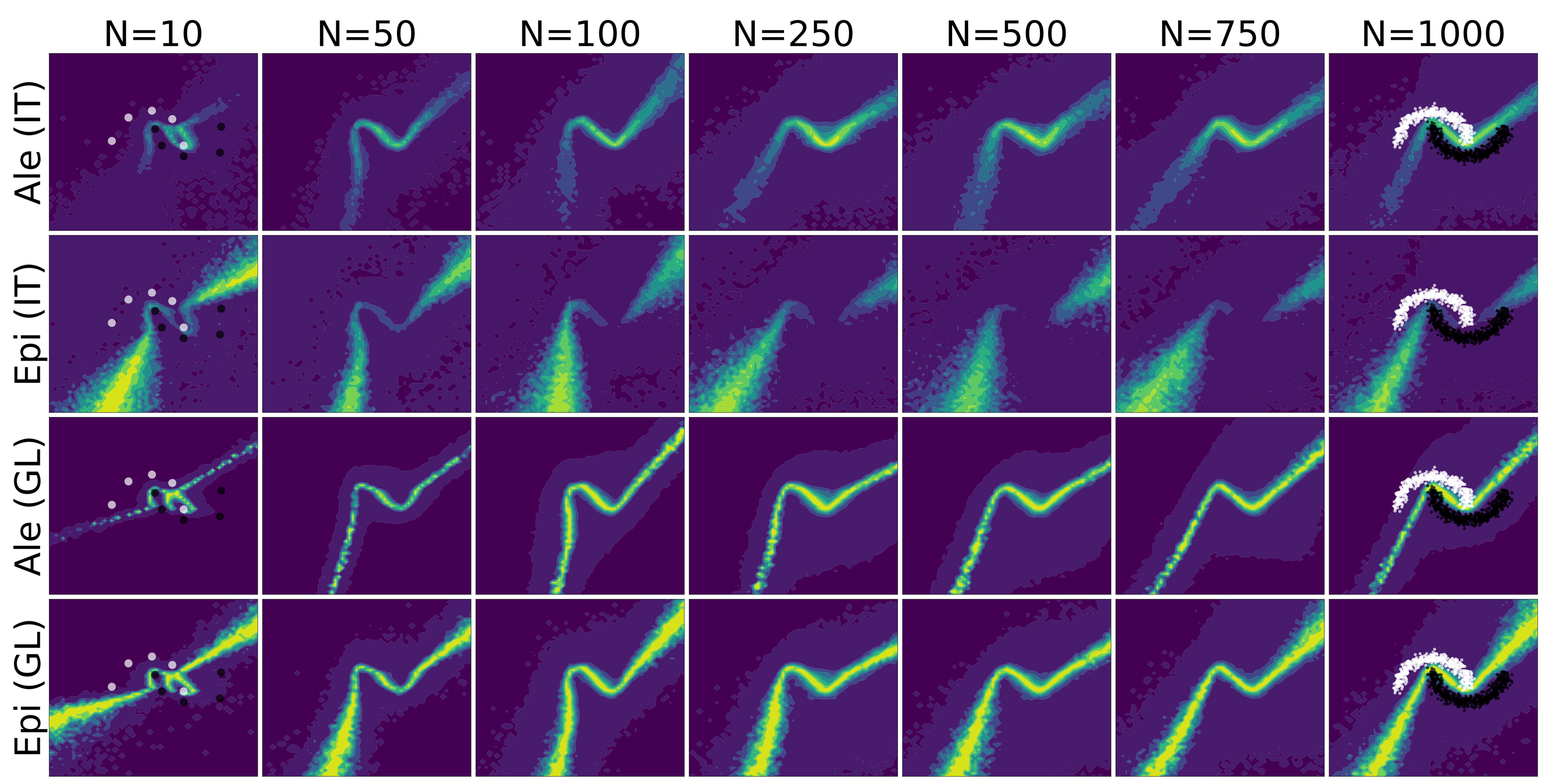
GL disentanglement [3] represents AU with variance in the logits. With a BNN we can use the **variance of the mean** to make predictions with EU. **Is this better than IT disentanglement?**

Experiment 1. EU should reduce with additional training data. AU should stay the same.



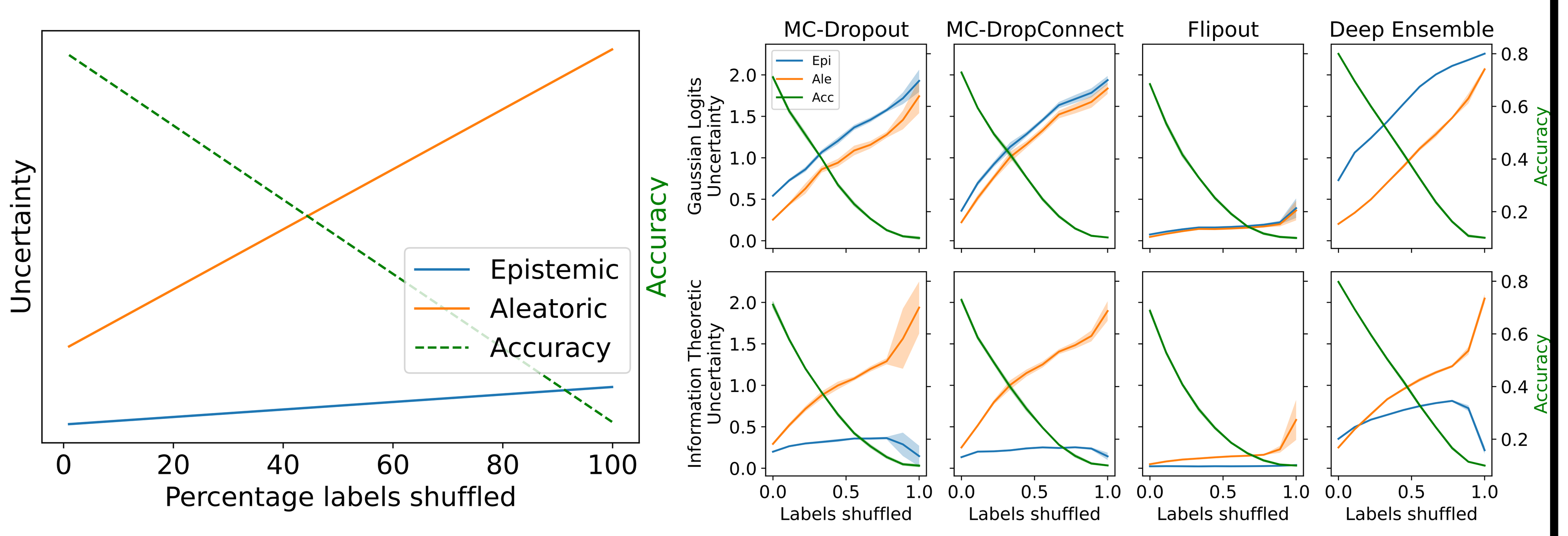
(a) (b)

Expectation (a) vs. Reality (b) of changing dataset size. AU increases for larger datasets?!



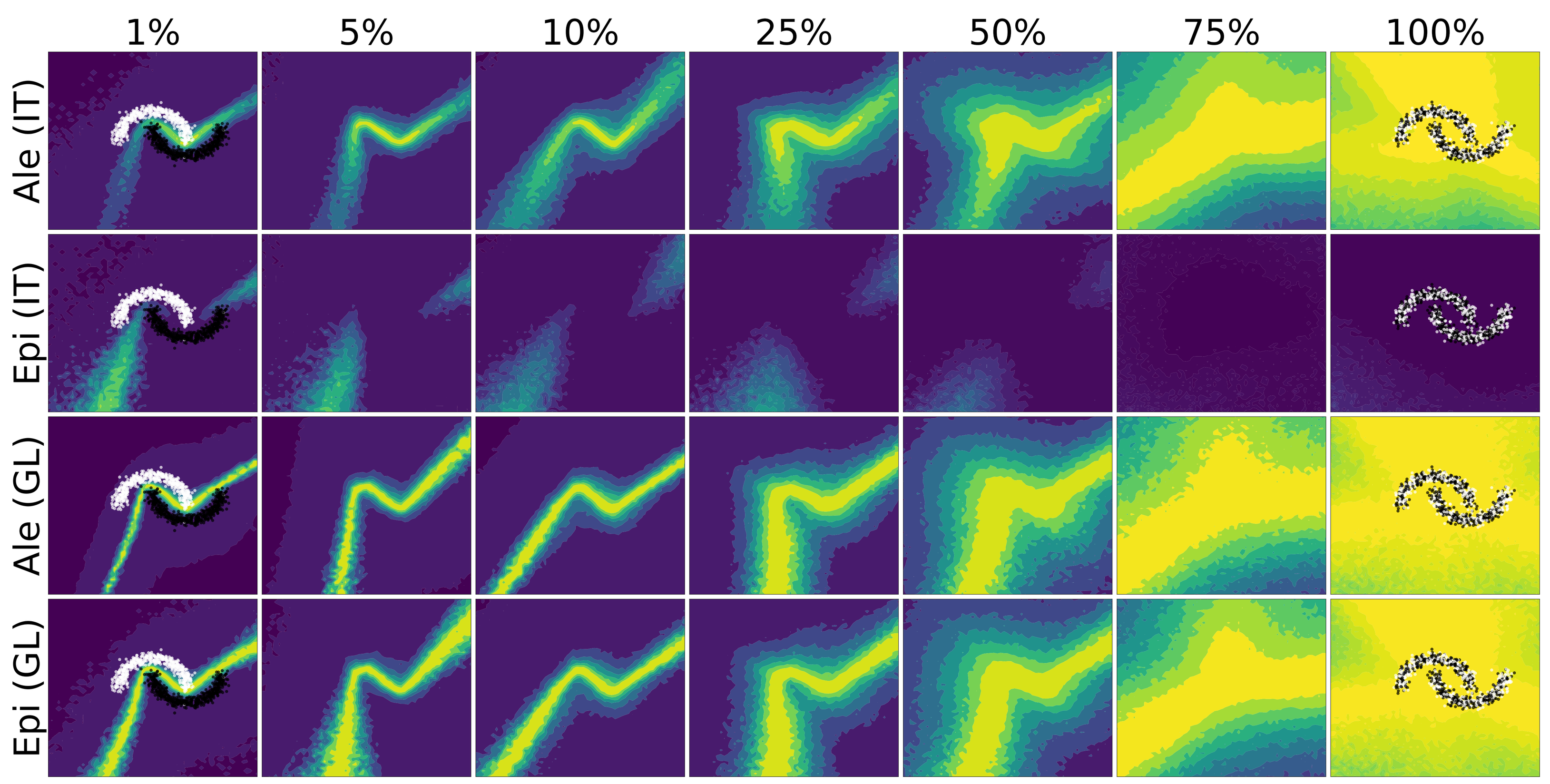
Toy data - increasing dataset size. IT exchanges Epi for Ale. GL predicts similar for Ale and Epi.

Experiment 2. AU should increase when labels are random. EU should stay similar.



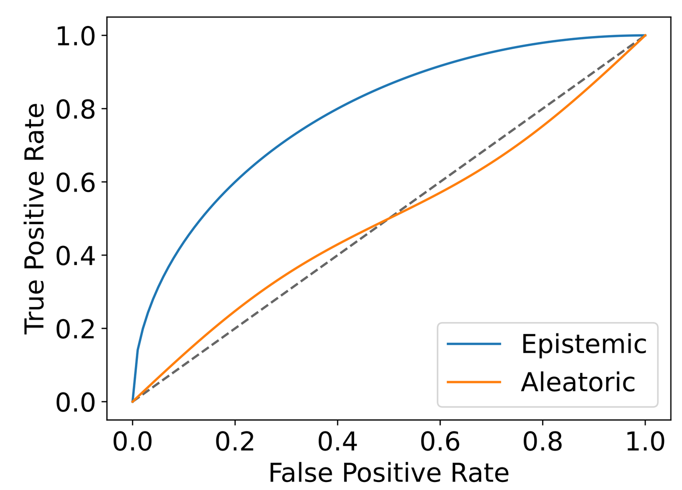
(a) (b)

Expectation (a) vs. Reality (b) for adding label noise. With GL the EU increases a lot with noisy lables?!



Toy data - increasing label noise. IT: Epi decreases. GL: Epi follows Ale.

Experiment 3. EU should be high when samples are out-of-distribution (OoD). AU should be random.



Expectation:
OoD samples → high EU → good ROC

Reality: Good ROC-AUC on OOD detection for AU and EU?!

CIFAR10, $\sigma_{\bar{x}} \approx .002$	GL AU	GL EU	IT AU	IT EU
MC-Dropout	0.644	0.642	**0.651**	0.649
MC-DropConnect	0.650	0.657	0.657	**0.658**
Flipout	0.626	**0.629**	0.625	0.579
Deep Ensembles	0.679	**0.709**	0.689	0.701
FashionMNIST, $\sigma_{\bar{x}} \approx .003$	GL AU	**GL EU**	IT AU	IT EU
MC-Dropout	0.753	**0.769**	0.761	0.764
MC-DropConnect	0.748	**0.780**	0.766	0.746
Flipout	0.649	**0.673**	0.661	0.579
Deep Ensembles	0.768	**0.811**	0.780	0.787
Wine, $\sigma_{\bar{x}} \approx .010$	**GL AU**	**GL EU**	**IT AU**	IT EU
MC-Dropout	**0.971**	0.961	0.943	0.670
MC-DropConnect	**0.959**	**0.957**	**0.954**	0.883
Flipout	**0.981**	**0.981**	**0.982**	0.974
Deep Ensembles	**0.985**	**0.984**	**0.981**	0.952

Takeaways

1. **We cannot separate aleatoric and epistemic uncertainty.**
2. GL EU is good for OoD detection because it includes AU.

References

[1] Lewis Smith and Yarin Gal. Understanding measures of uncertainty for adversarial example detection. *Uncertainty in Artificial Intelligence*, 2018.
[2] Lisa Wimmer et al. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? *Uncertainty in Artificial Intelligence*, 2023.
[3] Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? *NIPS*, 2017.