Weekly Study Report

Wang Ma

2025-03-11

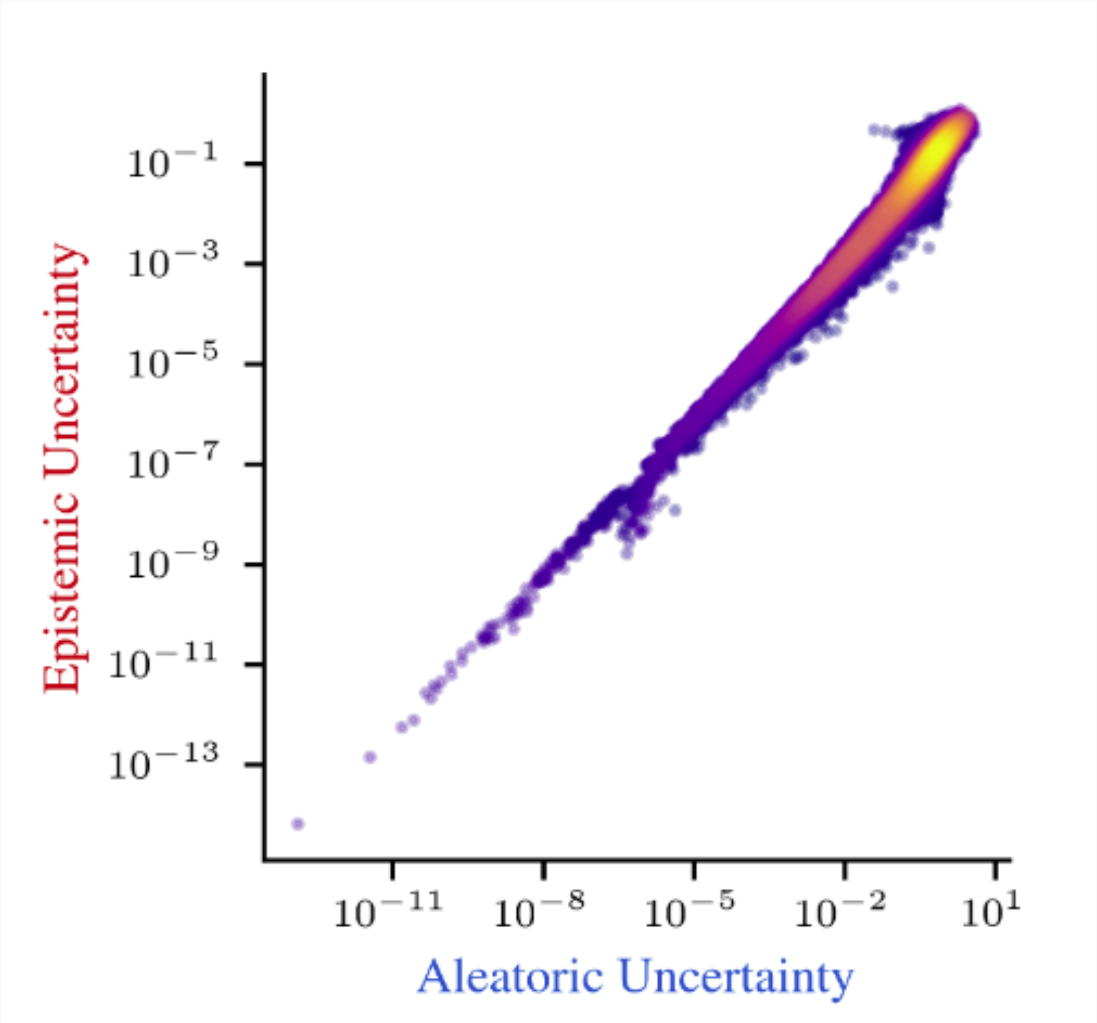
Electrical, Computer, and Systems Engineering Department
Rensselaer Polytechnic Institute

Outline

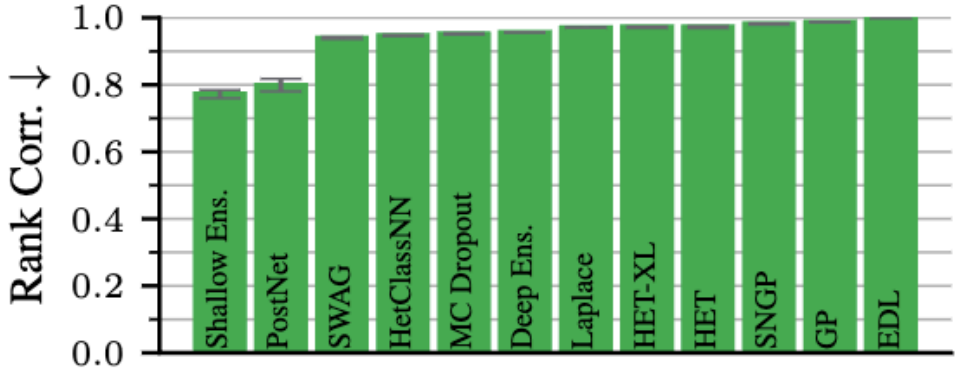
1. (NeurIPS 2024) Benchmarking Uncertainty Disentanglement 2
2. Ensemble Size Analysis 7
3. Credal Set: Which Point to Predict? 16

1. (NeurIPS 2024) Benchmarking Uncertainty Disentanglement

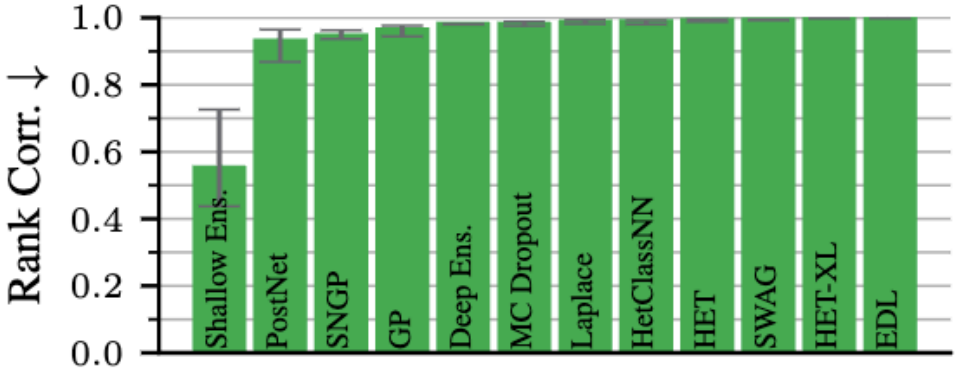
1.1 High Correlation of AU and EU



1.1 High Correlation of AU and EU



(a) ImageNet results. All twelve distributional methods exhibit a high rank corr. (≥ 0.78).

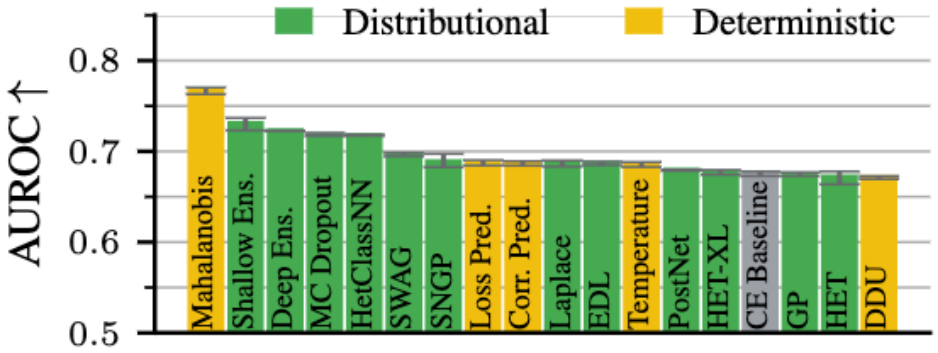


(b) CIFAR-10 results. Eleven out of twelve distributional methods exhibit a strong rank corr. (≥ 0.93).

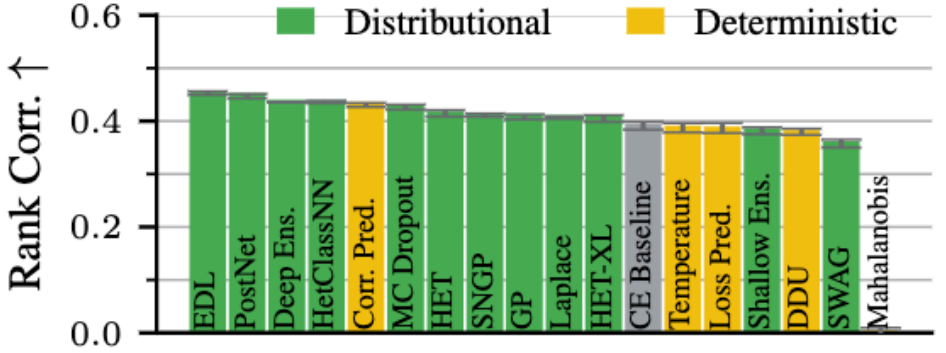
Figure 2: Rank correlation between the aleatoric and epistemic estimates obtained by the IT decomposition on ImageNet (left) and CIFAR-10 (right). The two uncertainty components are strongly correlated for most methods, violating a necessary condition of their disentanglement.

1.2 Evaluating EU and AU

- EU: Set corrupted ImageNet as OOD, and do OOD detection !
- AU: Using Multi-Label Dataset, Ground Truth AU = $\mathbb{H}$(Multi Label Distribution) ?



(a) OOD detection AUROC results. OOD samples are perturbed by ImageNet-C corruptions of severity two. Mahalanobis, the best method, is trained specifically to distinguish OOD data of this severity.



(b) Rank correlation of uncertainty estimators and the GT aleatoric uncertainty on ImageNet. The entropy of the ImageNet-ReaL label distributions is used as GT aleatoric uncertainty.

Figure 3: Performance of uncertainty quantification methods on epistemic (left) and aleatoric (right) uncertainty tasks on the ImageNet validation dataset.

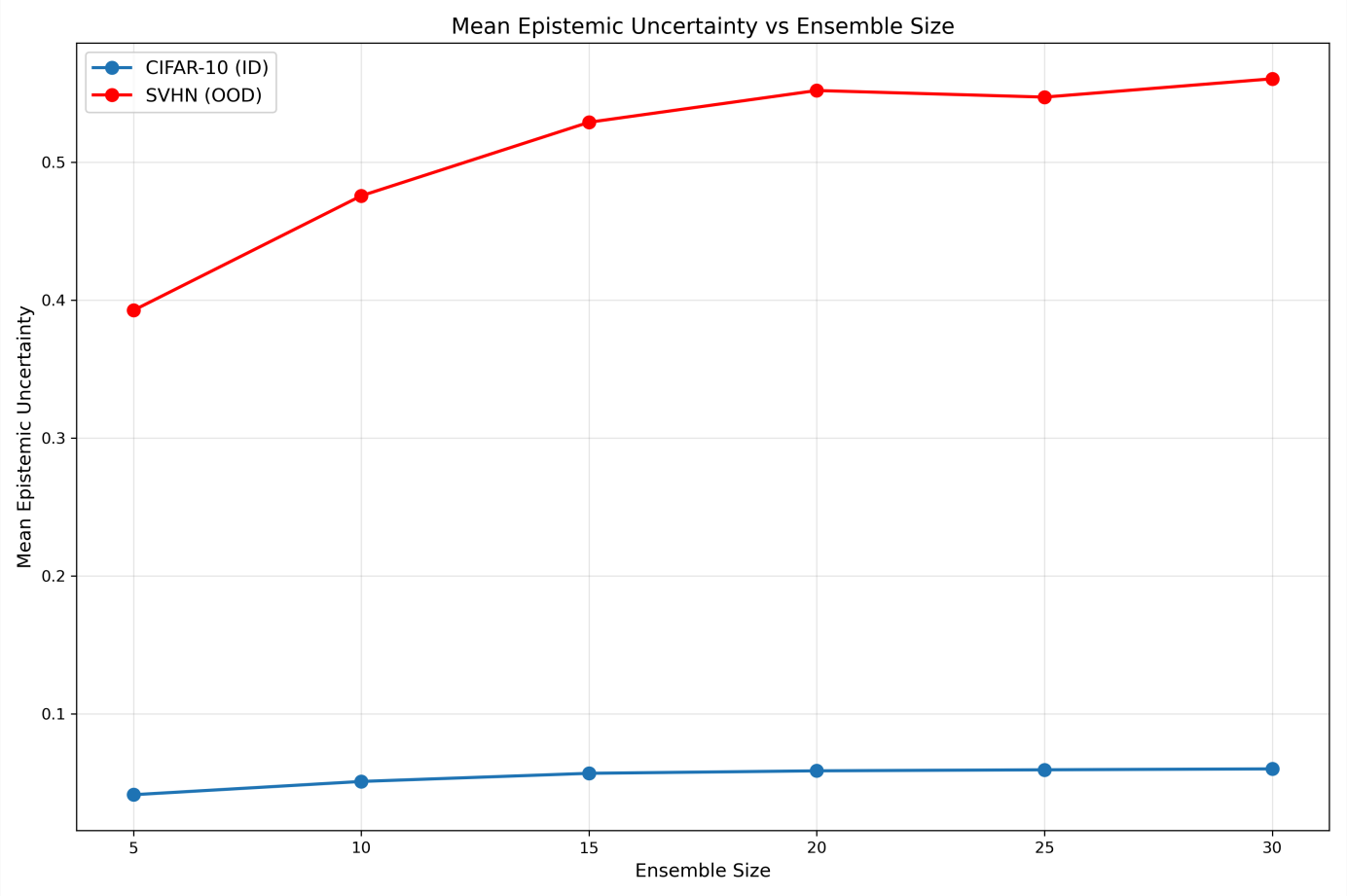
1.3 Other Findings

1. **Different tasks require different methods.** Method performs well on one task does not generalize well on other tasks.
2. **Uncertainties are robust on OOD data (domain shift).** This means even the model performance (accuracy) decreases significantly on OOD data, but the quantified uncertainties are robust and reliable.
3. **CIFAR-10 results do not always transfer to ImageNet.**
   - Conclusion on different scales of dataset are not unified.
   - UQ on ImageNet is more robust on CIFAR-10. On Corrupted ImageNet, model consistently shows high uncertainty on wrongly-classified samples, while on CIFAR-10, model does not have this kind of robustness.

2. Ensemble Size Analysis

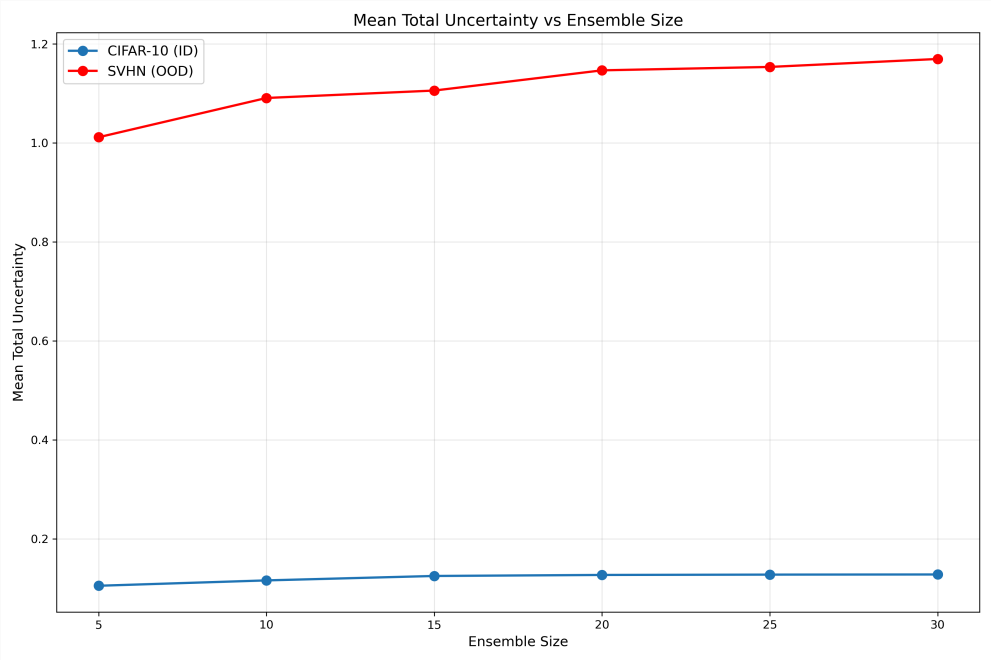
2.1 Increasing Uncertainty as Ensemble Size Increases

Epistemic Uncertainty on ID data and OOD data

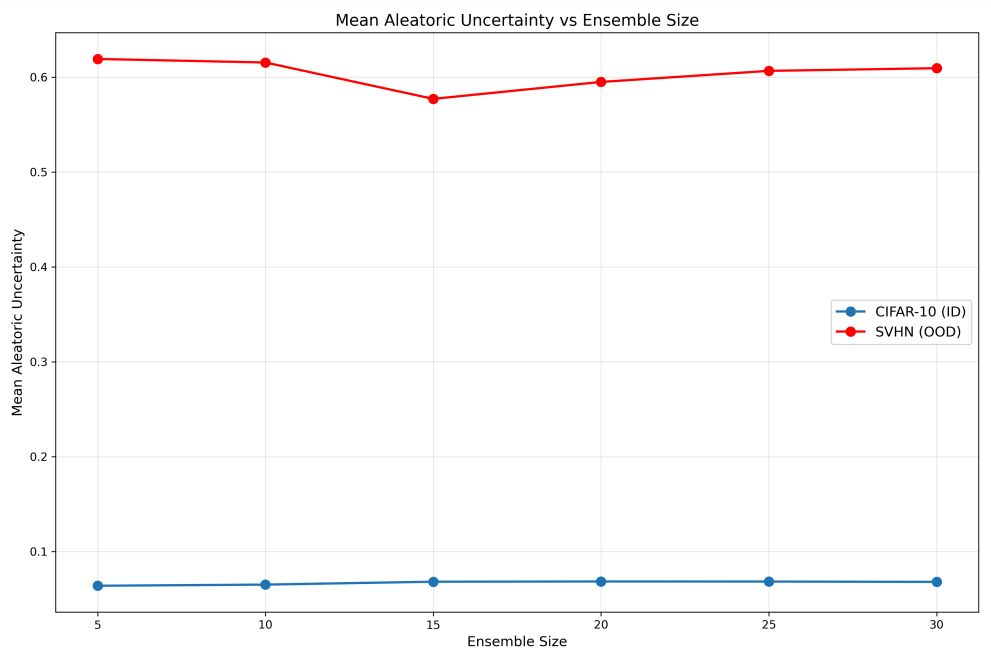


2.1 Increasing Uncertainty as Ensemble Size Increases

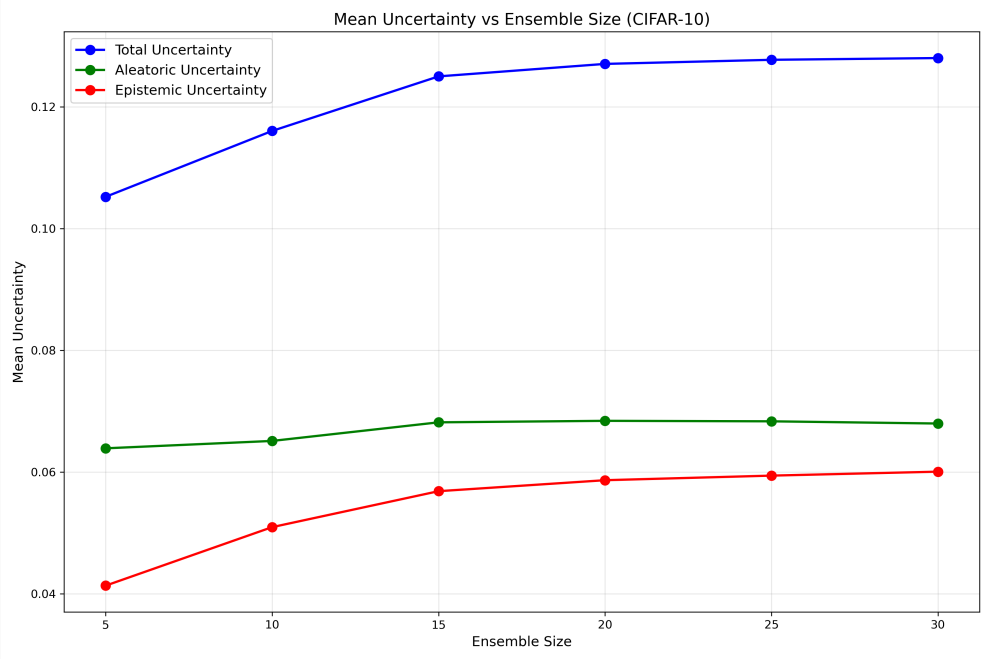
Total Uncertainty

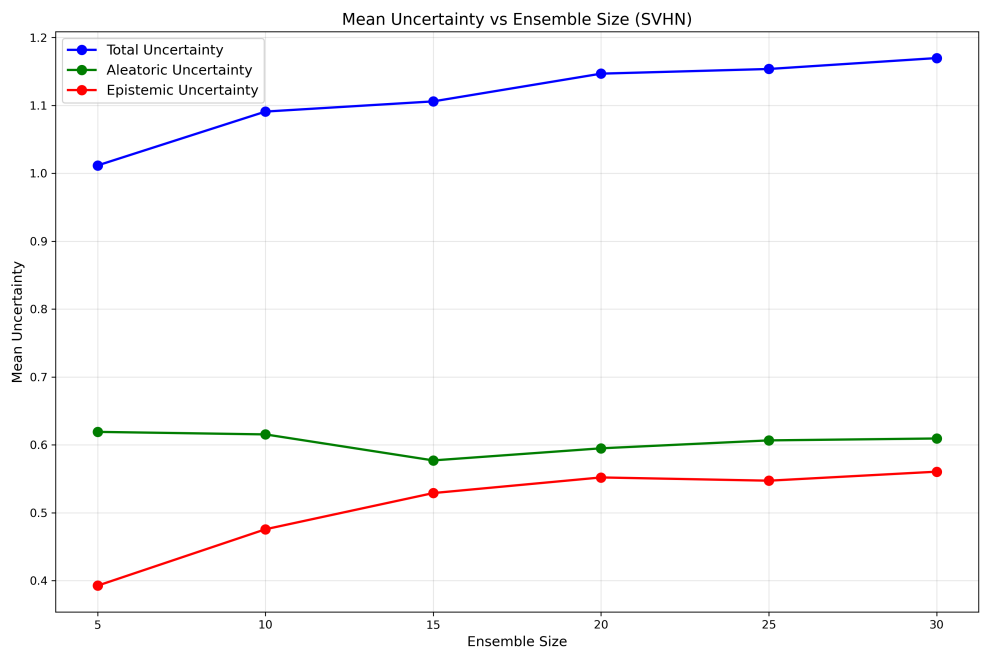


Aleatoric Uncertainty



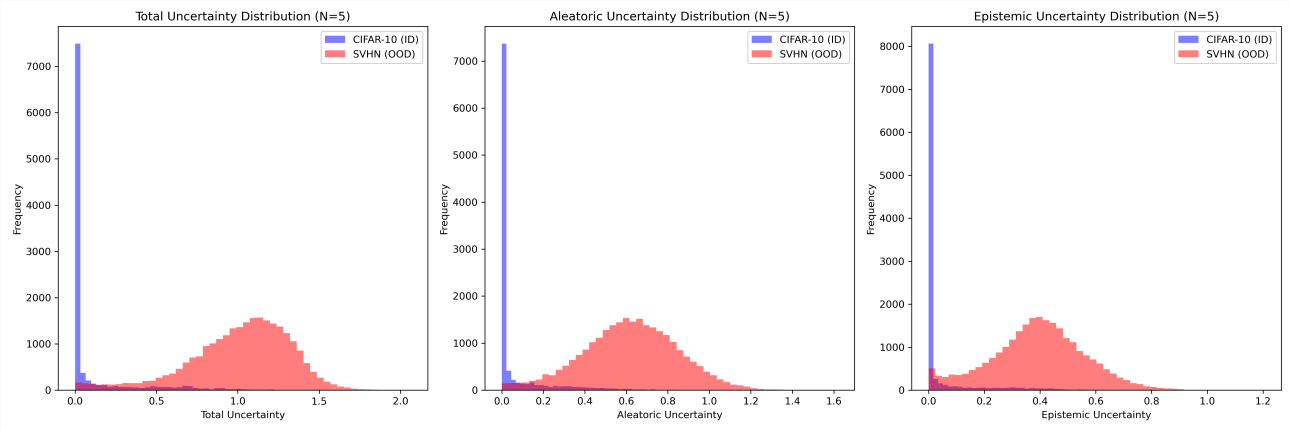
2.1 Increasing Uncertainty as Ensemble Size Increases

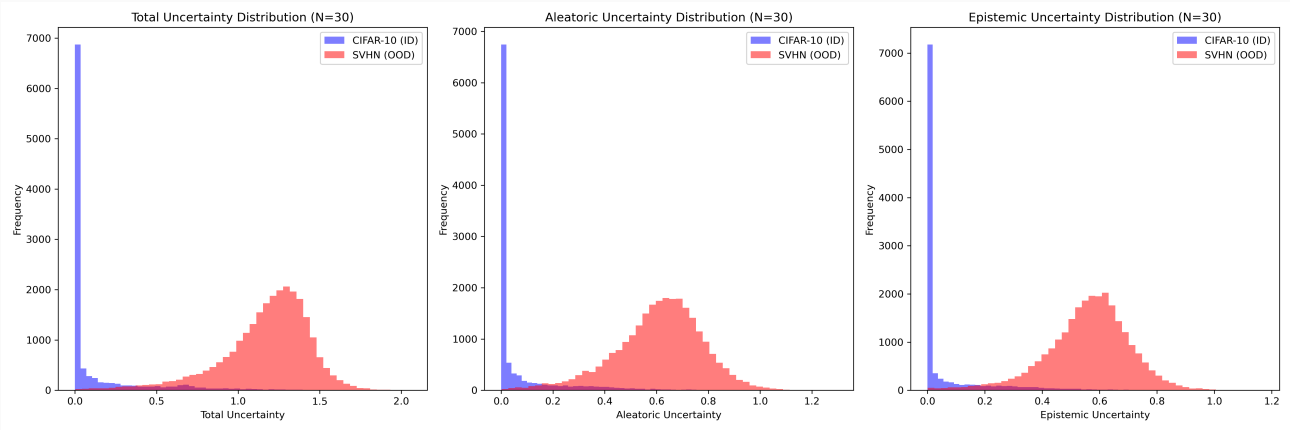




2.2 Uncertainty Distribution with Different Ensemble Size

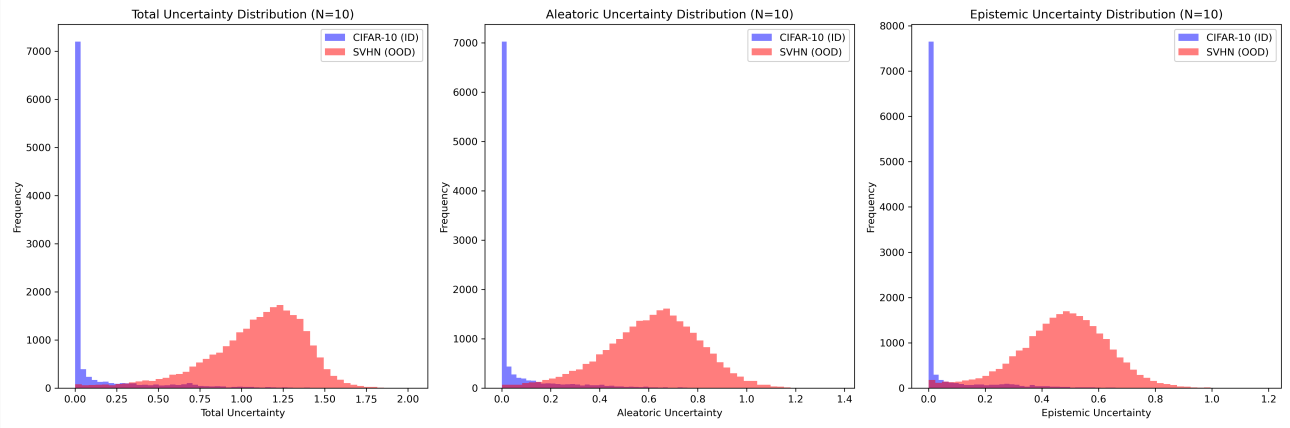
Ensemble of 5 and Ensemble of 30

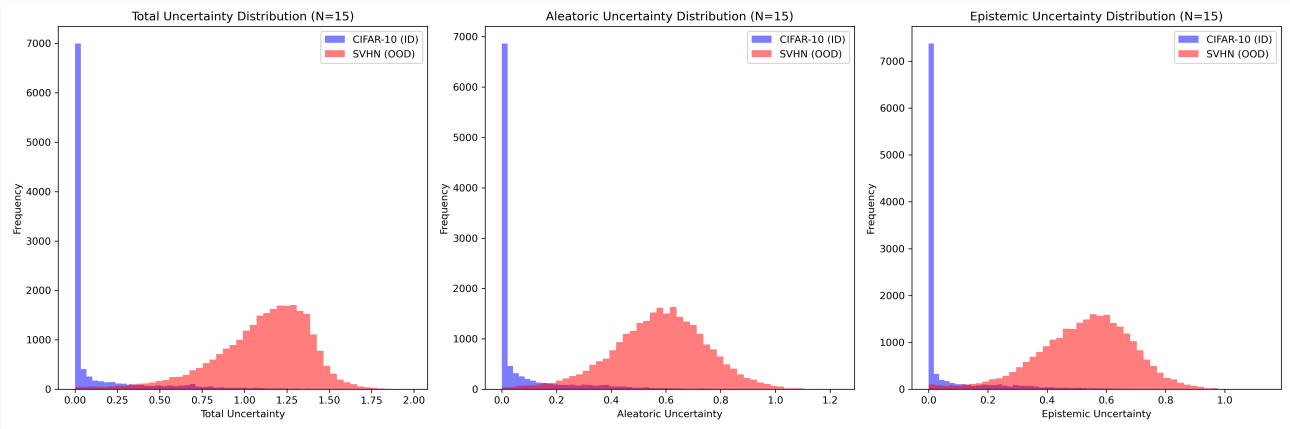




2.2 Uncertainty Distribution with Different Ensemble Size

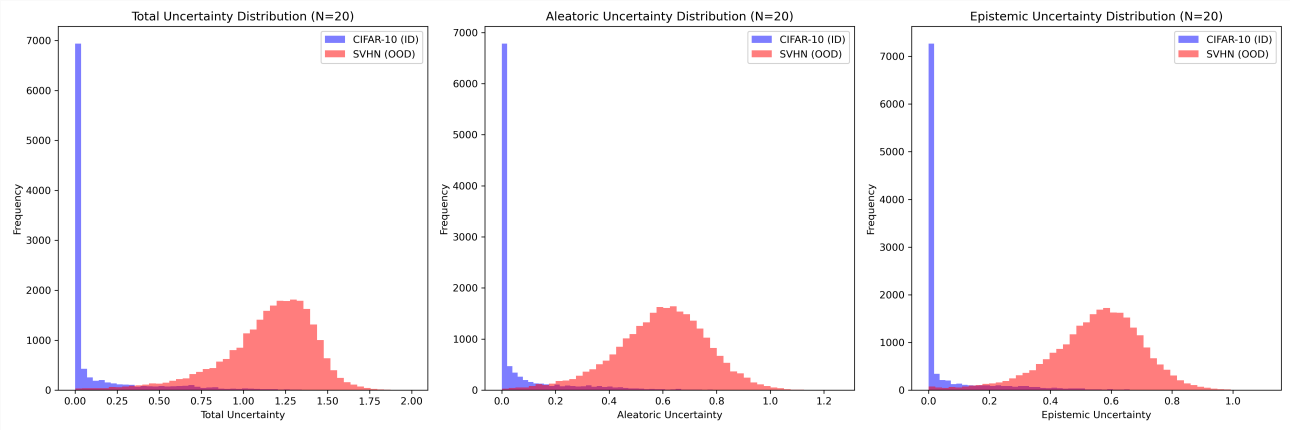
Ensemble of 10 and Ensemble of 15

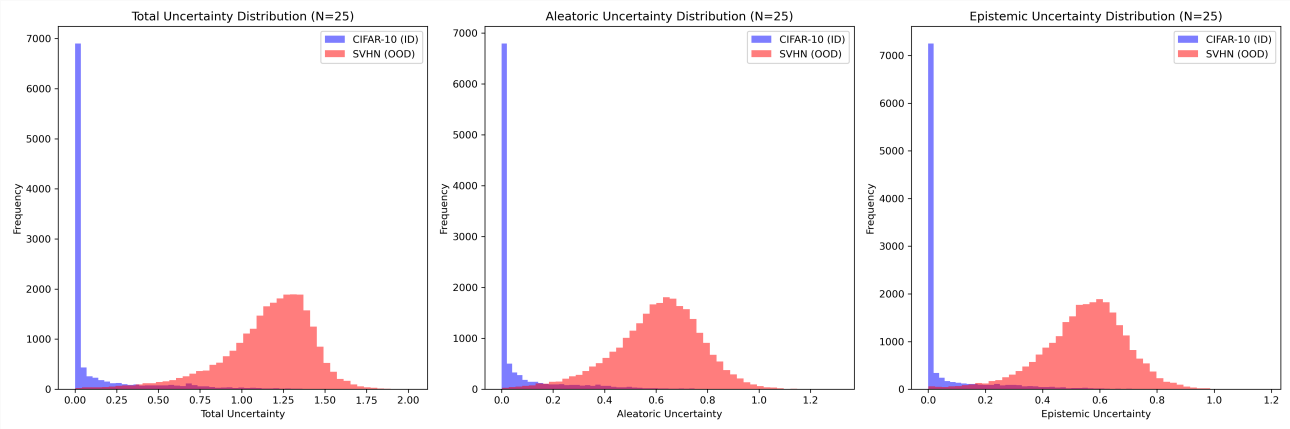




2.2 Uncertainty Distribution with Different Ensemble Size

Ensemble of 20 and Ensemble of 25





2.2 Uncertainty Distribution with Different Ensemble Size

A Summary of Uncertainties on In-Distribution Data

	ID: Epistemic Uncertainty					ID: Aleatoric Uncertainty					ID: Total Uncertainty				
	Mean	Median	Std	Min	Max	Mean	Median	Std	Min	Max	Mean	Median	Std	Min	Max
Ensemble Size of 5	0.041320	0.000071	0.110122	0.000000	0.864338	0.063883	0.000793	0.146188	0.000000	1.122580	0.105203	0.000868	0.243199	0.000000	1.645585
Ensemble Size of 10	0.050931	0.000176	0.121101	0.000000	0.891909	0.065096	0.001213	0.141068	0.000000	0.997535	0.116027	0.001408	0.253785	0.000000	1.679240
Ensemble Size of 15	0.056839	0.000305	0.127462	0.000000	0.886922	0.068152	0.001713	0.142471	0.000000	1.035193	0.124992	0.002037	0.263489	0.000000	1.695852
Ensemble Size of 20	0.058643	0.000378	0.128247	0.000000	0.917728	0.068399	0.001818	0.141857	0.000000	0.982105	0.127042	0.002214	0.264727	0.000000	1.772659
Ensemble Size of 25	0.059397	0.000427	0.128690	0.000000	0.962166	0.068319	0.001904	0.141368	0.000000	1.012856	0.127716	0.002371	0.265409	0.000000	1.750346
Ensemble Size of 30	0.060040	0.000504	129.000000	0.000000	0.923000	0.067900	0.002030	0.140000	0.000000	0.992400	0.128000	0.002560	0.265000	0.000000	1.744000

A Summary of Uncertainties on Out-of-Distribution Data

	OOD : Epistemic Uncertainty					OOD : Aleatoric Uncertainty					OOD: Total Uncertainty				
	Mean	Median	Std	Min	Max	Mean	Median	Std	Min	Max	Mean	Median	Std	Min	Max
Ensemble Size of 5	0.392536	0.396798	0.169551	0.000056	1.205149	0.619065	0.623158	0.229389	0.000643	1.618425	1.011601	1.056091	0.314094	0.000722	2.064153
Ensemble Size of 10	0.475581	0.483683	0.159887	0.000105	1.185373	0.615351	0.629631	0.196488	0.001056	1.372019	1.090932	1.140307	0.297090	0.001160	2.018377
Ensemble Size of 15	0.528920	0.542861	0.160365	0.000189	1.133958	0.577018	0.584516	0.173846	0.001328	1.227538	1.105938	1.152126	0.283258	0.001621	1.981715
Ensemble Size of 20	0.551943	0.566647	0.153212	0.000248	1.103491	0.594856	0.606186	0.172679	0.001112	1.251327	1.146799	1.194277	0.278577	0.001361	1.996025
Ensemble Size of 25	0.547154	0.560456	0.148679	0.000367	1.174035	0.606526	0.622013	0.170917	0.001591	1.304528	1.153680	1.202047	0.276981	0.001958	2.015223
Ensemble Size of 30	0.560000	0.572000	0.145000	0.000459	1.169000	0.609000	0.624000	0.168000	0.001810	1.296000	1.169000	1.218000	0.274000	0.002290	2.055900

So, the ensemble size will many influence Epistemic Uncertainty! (Mode models, more disagreements!) The Size of ensemble model almost does no change AU.

2.3 Comparison on OOD Detection

Do OOD detection with different ensemble sizes.

- ID: CIFAR-10
- OOD: SVHN

Here the results of "Ensemble size of 5" is the average of 5 trials (randomly choosing 5 models from the 30 models pool)

	AUROC	AUPR
Ensemble Size of 30	**0.9779**	**0.9888**
Ensemble Size of 5	0.9599	0.9772

3. Credal Set: Which Point to Predict?

3. Credal Set: Which Point to Predict?

	Cifar 10 (Resnet 18)	Cifar 100 (Resnet 50)
Deep Ensemble	95.95%	85.35%
Credal Wrapper (Mid Point)	95.91%	84.93%
Credal Wrapper (Lowest Entropy Point)	95.91%	84.92%
Credal Wrapper (Highest Entropy Point)	93.36%	79.23%
Credal Wrapper (Lower Bound)	95.89%	84.84%
Credal Wrapper (Upper Bound)	95.98%	84.76%