NEURIPS 2023 SPOTLIGHT
NEW ORLEANS, USA

IMPLICIT VARIATIONAL INFERENCE FOR HIGH-DIMENSIONAL POSTERIORS

Anshuk Uppal*, Kristoffer Stensbo-Smidt*, Wouter Boomsma°, Jes Frellsen*

* = Technical University of Denmark, ° = University of Copenhagen

VARIATIONAL INFERENCE FOR DEEP MODELS

A. Bayesian inference can provide excellent model generalisation and calibration to deep overparameterised models.

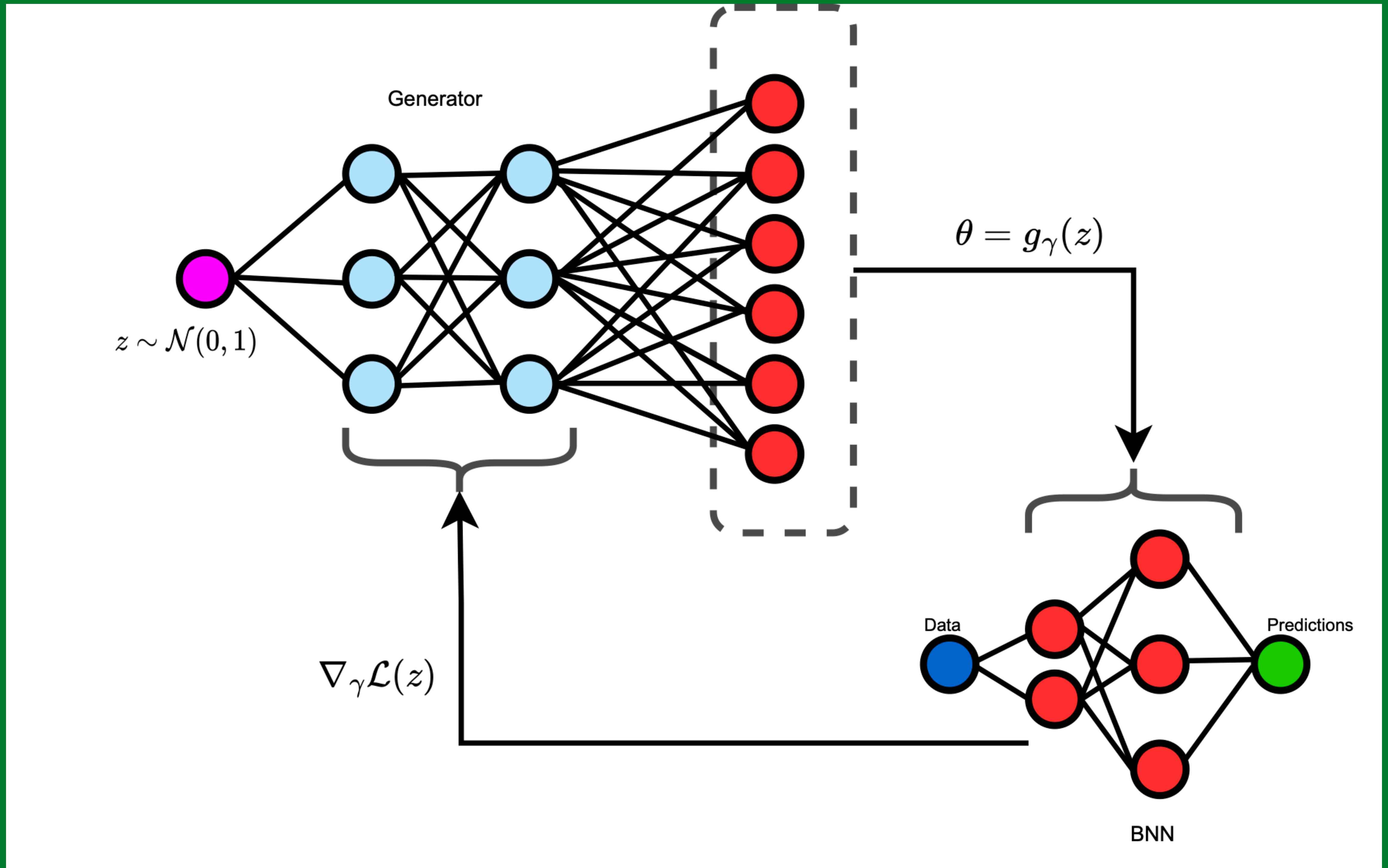
B. Accurate Bayesian inference is impossible for any complex NN model and hence in practice we rely on approximate inference.

C. Variational inference optimises over a chosen family of distributions to approach the true posterior.

D. The efficacy of VI hinges on this choice, we propose to use a highly flexible family of distributions called implicit distributions.

IMPLICIT DISTRIBUTIONS

- Easy to sample from but have no closed form density for computing log-likelihoods.
- We call these networks generators or neural samplers. (Hypernetworks)



CHALLENGES

1. For VI we need to measure a KL which is not trivial with implicit distributions.

DEF^n VARIATIONAL APPROX.

BASE DIST

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \int q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z}) q(\boldsymbol{z}) \, \mathrm{d}\boldsymbol{z} = \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})}[q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z})], \quad \text{where,}$$

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{\theta} \mid g_{\boldsymbol{\gamma}}(\boldsymbol{z}), \sigma^2 \boldsymbol{I}_m), \quad g_{\boldsymbol{\gamma}} : \mathbb{R}^d \to \mathbb{R}^m,$$

Generator

2. Cannot evaluate entropy of an implicit distribution.

ELBO: $$\mathcal{L}(\gamma) = \mathbb{E}_{q_\gamma(\theta)}\left[\log p(\theta, \mathcal{D})\right] - \mathbb{E}_{q_\gamma(\theta)}\left[\log q_\gamma(\theta)\right]$$

Entropy

3. Cannot evaluate gradients (w.r.t γ) of this unavailable entropy.

Footnote, add citations for UIVI, SIVI, D-SIVI

APPROXIMATING ENTROPY VIA LINEARISATION

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \int q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z})\, q(\boldsymbol{z})\, \mathrm{d}\boldsymbol{z} = \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})}[q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z})],$$

Non-conjugate

$$q_{\gamma}(\boldsymbol{\theta} \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{\theta} \mid g_{\boldsymbol{\gamma}}(\boldsymbol{z}), \sigma^2 \boldsymbol{I}_m), \quad g_{\boldsymbol{\gamma}} : \mathbb{R}^d \to \mathbb{R}^m,$$

Non-linear

- Linearise the g about its input using Taylor series $\longrightarrow$ $g_{\boldsymbol{\gamma}}(\boldsymbol{z}) \approx g_{\boldsymbol{\gamma}}(\boldsymbol{z}') + \boldsymbol{J}_g(\boldsymbol{z}')\,(\boldsymbol{z} - \boldsymbol{z}') := T^1_{\boldsymbol{z}'}(\boldsymbol{z})$

$$q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z}) \approx \tilde{q}_{\boldsymbol{z}'}(\theta \mid \boldsymbol{z}) = \mathcal{N}(\boldsymbol{\theta} \mid T^1_{\boldsymbol{z}'}(\boldsymbol{z}), \sigma^2 \boldsymbol{I}_m)$$

TRACTABLE: $q_{\boldsymbol{\gamma}}(\boldsymbol{\theta}) = \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})}[q_{\boldsymbol{\gamma}}(\boldsymbol{\theta} \mid \boldsymbol{z})] \approx \mathbb{E}_{\boldsymbol{z} \sim q(\boldsymbol{z})}[\tilde{q}_{\boldsymbol{z}'}(\boldsymbol{\theta} \mid \boldsymbol{z})]$

APPROXIMATE ELBO AND SCALABILITY

APPROX. ENTROPY

$$H[q_{\boldsymbol{\gamma}}(\boldsymbol{\theta})] \approx \frac{1}{2}\mathbb{E}_{\boldsymbol{z}\sim q(\boldsymbol{z})}\left[\log\det\left(\boldsymbol{J}_g(\boldsymbol{z})\boldsymbol{J}_g(\boldsymbol{z})^\intercal + \sigma^2\boldsymbol{I}_m\right)\right] + \frac{m}{2} + \frac{m}{2}\log 2\pi$$

- Jacobians and their log determinant are very expensive, so can further derive a lower bound using fundamental LA theorems.

If $s_d(\boldsymbol{z}) \geq \ldots \geq s_1(\boldsymbol{z})$ are the non-zero singular values of the Jacobian $\boldsymbol{J}_g(\boldsymbol{z})$

$$\frac{1}{2}\log\det(\boldsymbol{J}_g(\boldsymbol{z})\boldsymbol{J}_g(\boldsymbol{z})^\intercal + \sigma^2\boldsymbol{I}_m) = \frac{1}{2}\sum_{i=1}^{d}\log(s_i^2(\boldsymbol{z}) + \sigma^2) + \frac{m-d}{2}\log\sigma^2$$

$$\frac{1}{2}\sum_{i=1}^{d}\log(s_i^2(\boldsymbol{z}) + \sigma^2) + \frac{m-d}{2}\log\sigma^2 \geq \boxed{\frac{d}{2}\log(s_1^2(\boldsymbol{z}) + \sigma^2) + \frac{m-d}{2}\log\sigma^2}$$

Scalable Entropy approximation

TESTING THE VARIATIONAL APPROXIMATION

- We test our variational approximation using deep BNNs as they contain millions of global latent variables.
- In UCI regression benchmarks we compare our posterior quality with HMC, and we compare within the two entropy approximations.

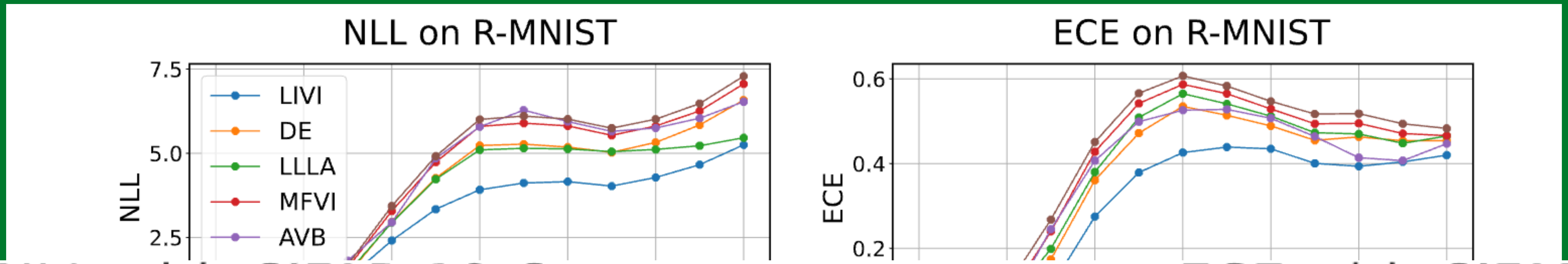
Table F.1: **UCI regression datasets.** We report RMSE (↓) on the test set and average across three different seeds for each model to quantify the variance in the results.

Method	Boston	Concrete	Energy	Kin8nm	Naval
LIVI ($\mathcal{L}'$)	2.32 ± 0.07	**4.24 ± 0.17**	0.41 ± 0.27	**0.03 ± 0.00**	**0.00 ± 0.00**
LIVI ($\mathcal{L}''$)	2.40 ± 0.09	4.62 ± 0.13	0.44 ± 0.11	0.08 ± 0.01	0.00 ± 0.01
HMC	**2.26 ± 0.00**	4.27 ± 0.00	**0.38 ± 0.00**	0.04 ± 0.00	**0.00 ± 0.00**
DE	3.28 ± 1.00	6.03 ± 0.58	2.09 ± 0.29	0.09 ± 0.00	0.00 ± 0.00
KIVI	2.80 ± 0.17	4.70 ± 0.12	0.47 ± 0.02	0.08 ± 0.00	0.00 ± 0.00
MNF	3.31 ± 0.10	5.82 ± 0.04	1.04 ± 0.01	0.08 ± 0.01	0.01 ± 0.00

Table F.2: **UCI regression datasets.** We report log-likelihood (↑) on the test set and average across three different seeds for each model to quantify the variance in the results.

Method	Boston	Concrete	Energy	Kin8nm	Naval
LIVI ($\mathcal{L}'$)	**−2.16 ± 0.05**	−2.79 ± 0.11	−1.17 ± 0.13	1.24 ± 0.04	6.74 ± 0.04
LIVI ($\mathcal{L}''$)	−2.40 ± 0.09	−2.99 ± 0.13	−1.37 ± 0.11	1.15 ± 0.01	5.84 ± 0.06
HMC	−2.20 ± 0.00	**−2.67 ± 0.00**	**−1.14 ± 0.00**	1.27 ± 0.00	**7.79 ± 0.00**
DE	−2.41 ± 0.25	−3.06 ± 0.18	−1.31 ± 0.22	**1.28 ± 0.02**	5.93 ± 0.05
KIVI	−2.53 ± 0.10	−3.05 ± 0.04	−1.30 ± 0.01	1.16 ± 0.01	5.50 ± 0.12
MNF	−2.66 ± 0.08	−3.24 ± 0.09	−1.34 ± 0.07	1.10 ± 0.01	5.01 ± 0.00

UQ TESTS ON MNIST & CIFAR10



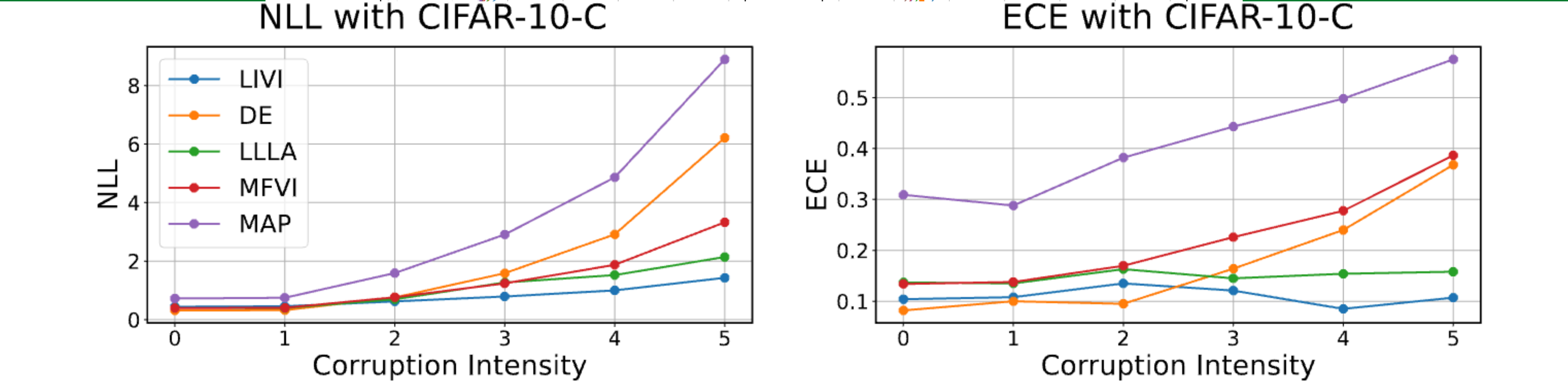
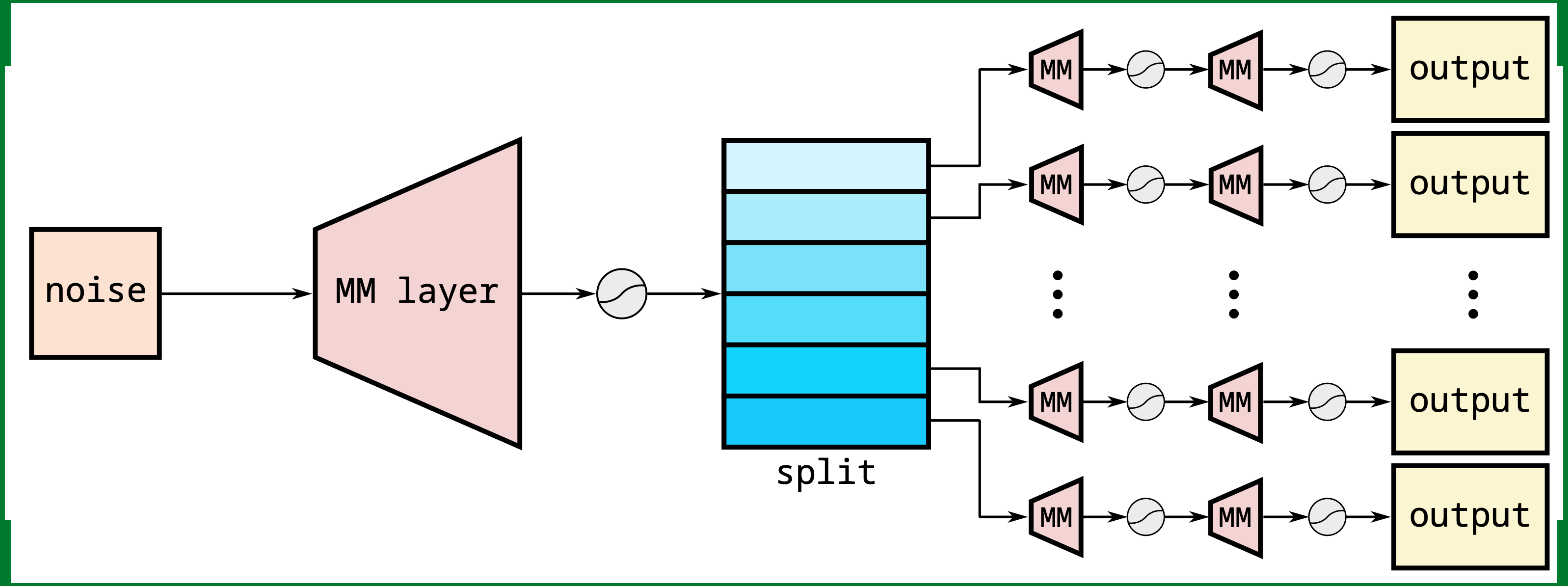


Figure 4: **OOD Test C2: Corrupted CIFAR10 benchmark.** OOD performance for methods trained on CIFAR10 and making predictions for CIFAR-10-C images corrupted with Gaussian blur (Hendrycks et al., 2019). LIVI performs as well or better than competitors.

SCALING TO 10s OF MILLIONS OF LATENT VARIABLES

- We also tested our approach on CIFAR100, using WideResNet(28,10), that contains roughly 36.5 million parameters.



CONCLUSION

- We present a novel entropy approximation to scale variational inference using implicit distributions.
- We lower bound this approximation further to make it computationally cheaper.
- We upgrade the MMNN* architecture to keep the number of generator parameters manageable.
- We outperform state of the art uncertainty quantification approaches while generating all the parameters of a BNN from a single generator, modelling within layer and across-layer parametric correlations.

*Kernel Implicit Variational Inference, Shi et. al. 2018