Approximating the Predictive Distribution via Adversarially-Trained Hypernetworks

Christian Henning[*,1], **Johannes von Oswald**[*,1], João Sacramento[1],
Simone Carlo Surace[2], Jean-Pascal Pfister[1,2] and Benjamin F. Grewe[1]

[1]Institute of Neuroinformatics, UZH / ETH Zurich, Switzerland
[2]Department of Physiology, University of Bern, Switzerland

Being able to model uncertainty is a vital property for any intelligent agent. Here we propose a novel approach for uncertainty estimation based on adversarially trained hypernetworks. We define a weight posterior to uniformly allow weight realizations of a neural network that meet a chosen fidelity constraint. This setting gives rise to a posterior predictive distribution that allows inference on unseen data samples. In this work, we train a combination of hypernetwork and main network via the GAN framework by sampling from this posterior predictive distribution. Due to the indirect training of the hypernetwork our method does not suffer from complicated loss formulations over weight configurations. We report empirical results that show that our method is able to capture uncertainty over outputs and exhibits performance that is on par with previous work. Furthermore, the use of hypernetworks allows producing arbitrarily complex, multi-modal weight posteriors.

Method

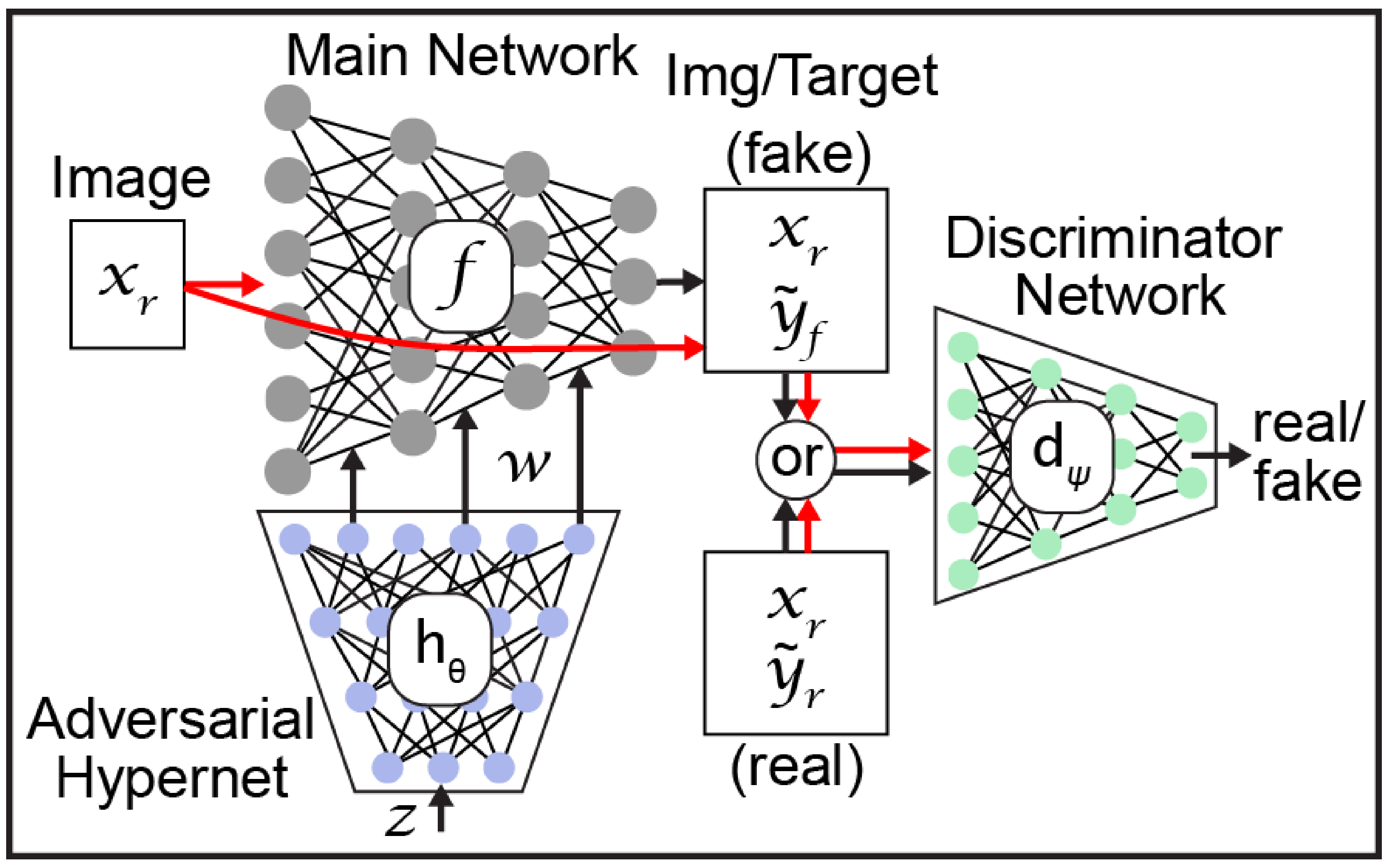

Figure 1: This figure illustrates our *Adversarial Hypernetwork*. The weights W of the main network $f(x, W)$ are generated by an auxiliary network h_θ. The discriminator d_ψ is used for training to ensure that h_θ generates weights that are appropriate for a desired output inference scheme $p(\tilde{y} \mid x, D)$.

In this work, we account for **weight uncertainty** to allow for an improved inference over the outputs of a neural network $f(x, W)$ with inputs x and weights W.

In the Bayesian view, given the posterior parameter distribution, such inference could be performed as follows:

$$\tilde{p}(\tilde{y} \mid x, D) = \int_W p(\tilde{y} \mid x, W)\, p(W \mid D)\, dW \tag{1}$$

We use a *hypernetwork*, that transforms $p_z(z)$ into $q(W)$, such that the inference scheme of eq. (1) is recovered:

$$q(\tilde{y} \mid x) = \int_W p(\tilde{y} \mid x, W)\, q(W)\, dW \tag{2}$$

We perform the matching of $\tilde{p}(\tilde{y} \mid x, D)$ and $q(\tilde{y} \mid x)$ by employing Least-Squares Generative Adversarial Networks [?].

Generation of Real Samples

As a heuristic, we consider a weight distribution defined through a *high-fidelity* constrained, e.g., classification accuracy: $\tau_f(W, D) = \frac{1}{N} \sum_n \mathbb{1}_{y_n = \arg\max_i f(x_n, W)_i}$:

$$\tilde{p}(W \mid D) = \frac{1}{Z_W} p_u(W)\, \Theta\left[\tau_f(W, D) - \tau^*\right] \tag{3}$$

Results

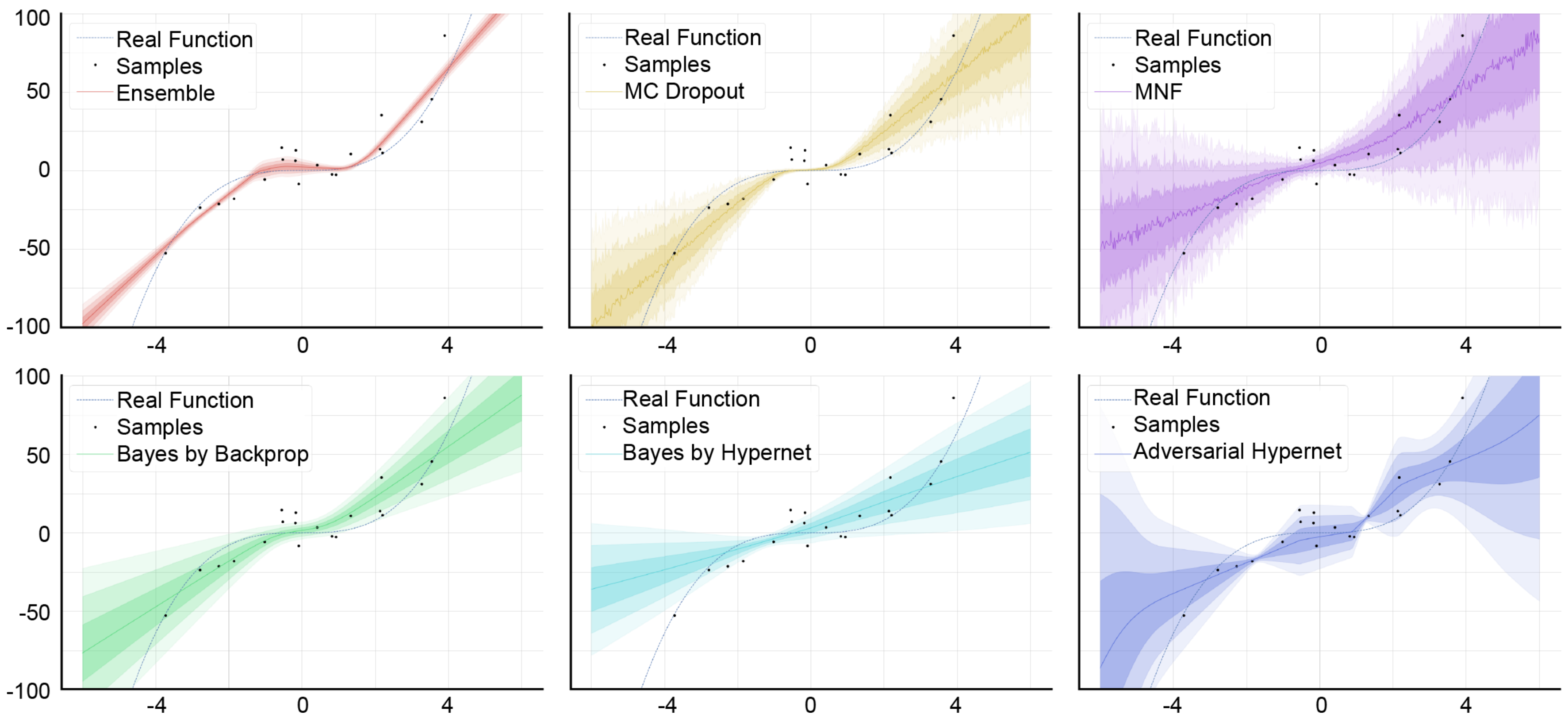

Figure 2: This figure depicts output uncertainty for a toy regression problem $g(x, \epsilon) = x^3 + \epsilon$, $\epsilon \sim \mathcal{N}(0, 9)$. Note, that all 20 training samples stem from the interval $x \in [-4, 4]$. The plots show the following methods (from top left to bottom right): Ensemble [?], MC Dropout [?], MNF [?], BbB [?], BbH [?] and the *Adversarial Hypernetwork* (ours).

On the toy regression illustrated in Figure 2, our approach shows a smooth approximation in areas that have a high density of training samples, while the uncertainty drastically increases in a non-linear fashion when training data points are sparse or absent.

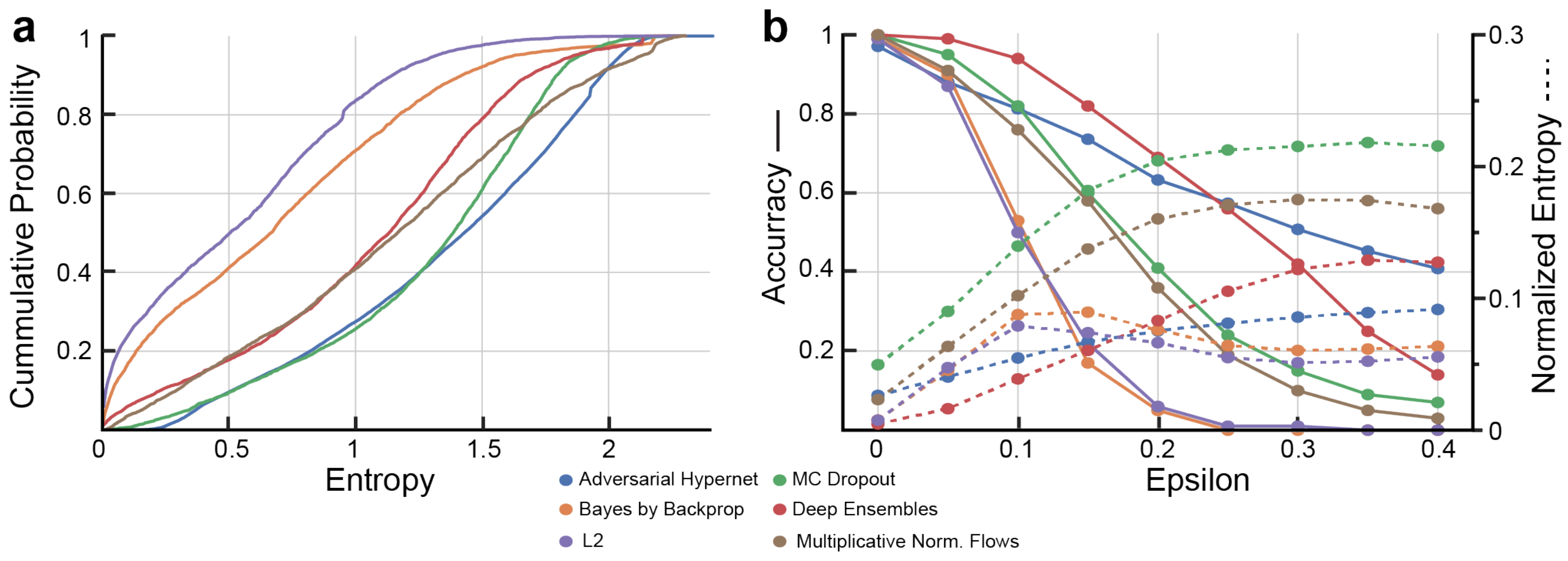

Figure 3: (a) Cumulative density function (CDF) of the entropies of the predictive distributions observed for notMNIST samples. (b) Evolution of accuracy (bold lines) and observed uncertainty (dashed lines) with increasing strength of an adversarial attack (FGSM [?]) on the inputs.

As shown in Figure 3 for a network trained on MNIST, our method successfully identifies unknown data through overall high prediction entropy and clearly performs on par with the state-of-the-art methods.

Conclusions

- Estimating a posterior predictive distribution that may account for weight uncertainty through a training process involving two auxiliary networks
  - Fast sampling of weight realizations through a *hypernetwork*
  - Distribution matching via an adversarial training scheme
- Output uncertainty estimations comparable to other state-of-the-art methods

References

Bayesian Deep Learning Workshop, Montréal, Canada, December 7, 2018

[*]These authors contributed equally to this work.