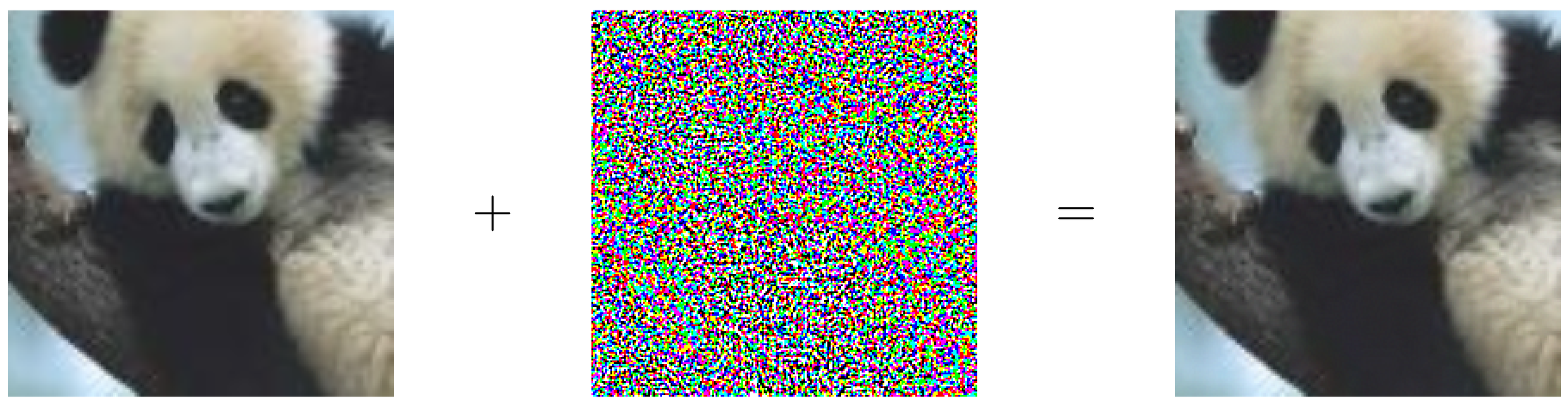
Bayesian Adversarial Spheres

Artur Bekasov, Iain Murray

{artur.bekasov, i.murray}@ed.ac.uk

Problem: Adversarial Examples



"panda" 57.7% confidence — *adversarial perturbation* — "gibbon" 99.3% confidence

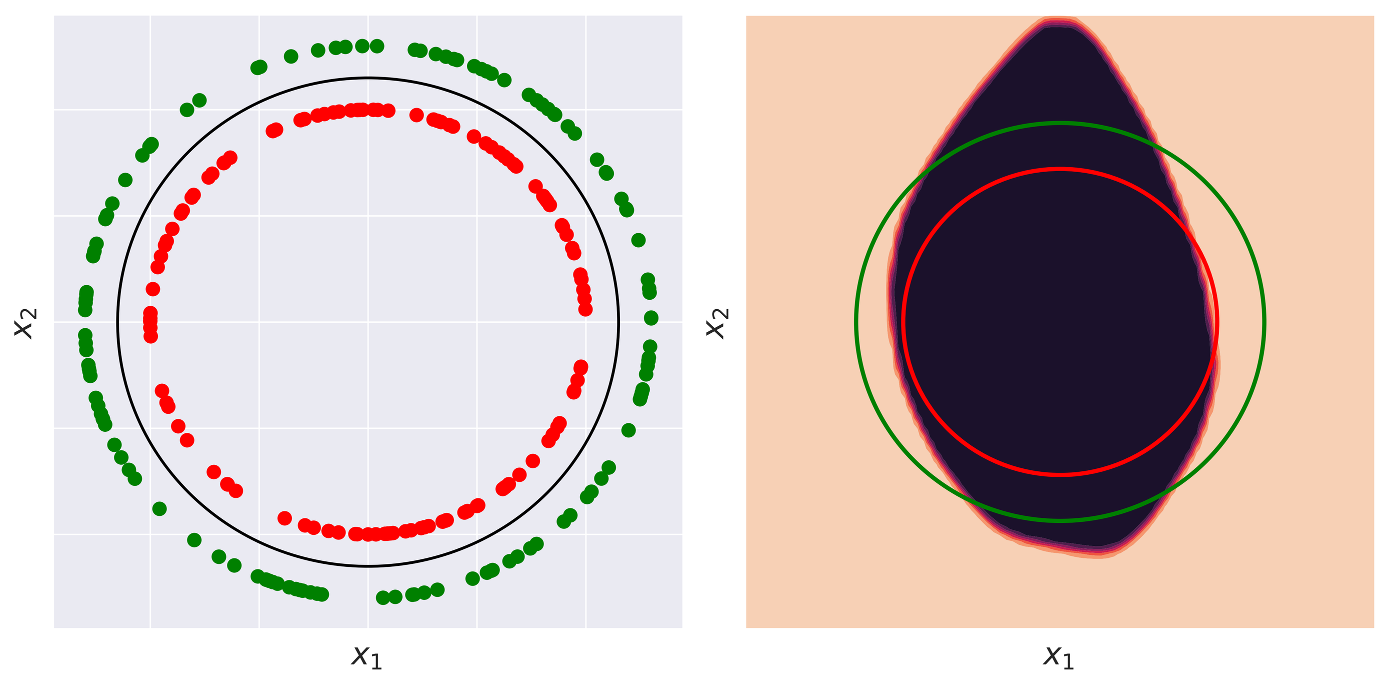
- Visually imperceptible changes in the image result in **confidently incorrect predictions**
- In practical decision making, a model should at least detect such changes and **become more uncertain** in its prediction

BNNs for detecting adv. examples

- It is difficult to cover a high-dimensional manifold with data. Regions exist where **different reasonable fits make different predictions**
- Capturing weight (*epistemic*) uncertainty $\Rightarrow$ better calibrated output uncertainty $\Rightarrow$ adversarial example detection
- *Bayesian neural networks (BNNs)* have been explored (Rawat et al. 2017; Bradshaw et al., 2017; Gal and Smith, 2018), but accurate posterior inference is difficult
- **Can we demonstrate the effect in a simpler setting, where inference is easier?**

Setup: Adversarial Spheres

- *Setup introduced by Gilmer et al. (2018)*



- A **binary classification** task using a synthetic dataset:

$$x^{(i)} = R\frac{z}{\|z\|_2} \qquad R = \begin{cases} 1, & \text{if } y^{(i)} = 0 \\ 1.3, & \text{if } y^{(i)} = 1 \end{cases} \qquad z \sim N(0, \mathbb{I})$$

- In a high-dimensional setting, Projected Gradient Descent finds adversarial examples **on the manifold** (the sphere surfaces), even for models with a **perfect validation score**

Model: Bayesian logistic regression

- **Logistic regression** with squared features:

$$p\left(y=1|x,w\right) = \sigma\left(w^T\phi(x)\right); \quad \phi(x) = \left[x_1^2 \; \ldots \; x_D^2\right]$$

- Represents axis-aligned ellipsoidal decision boundaries in D dimensions
- Inference still intractable, but **approx. inference is more accurate**

Hierarchical modelling

- Exploit symmetries using a hyper-prior on the mean:

$$w_i \sim N(\mu, \sigma_w^2); \quad \mu \sim N(0, \sigma_\mu^2)$$

- More expressive variational family using a hyper-prior on log-variance:

$$w_i \sim N(0, e^v); \quad v \sim N(0, \sigma_v^2)$$

- Hierarchical priors are useful for NN models as well (Neal, 1994). **How to choose them for real, complex problems?**

Results

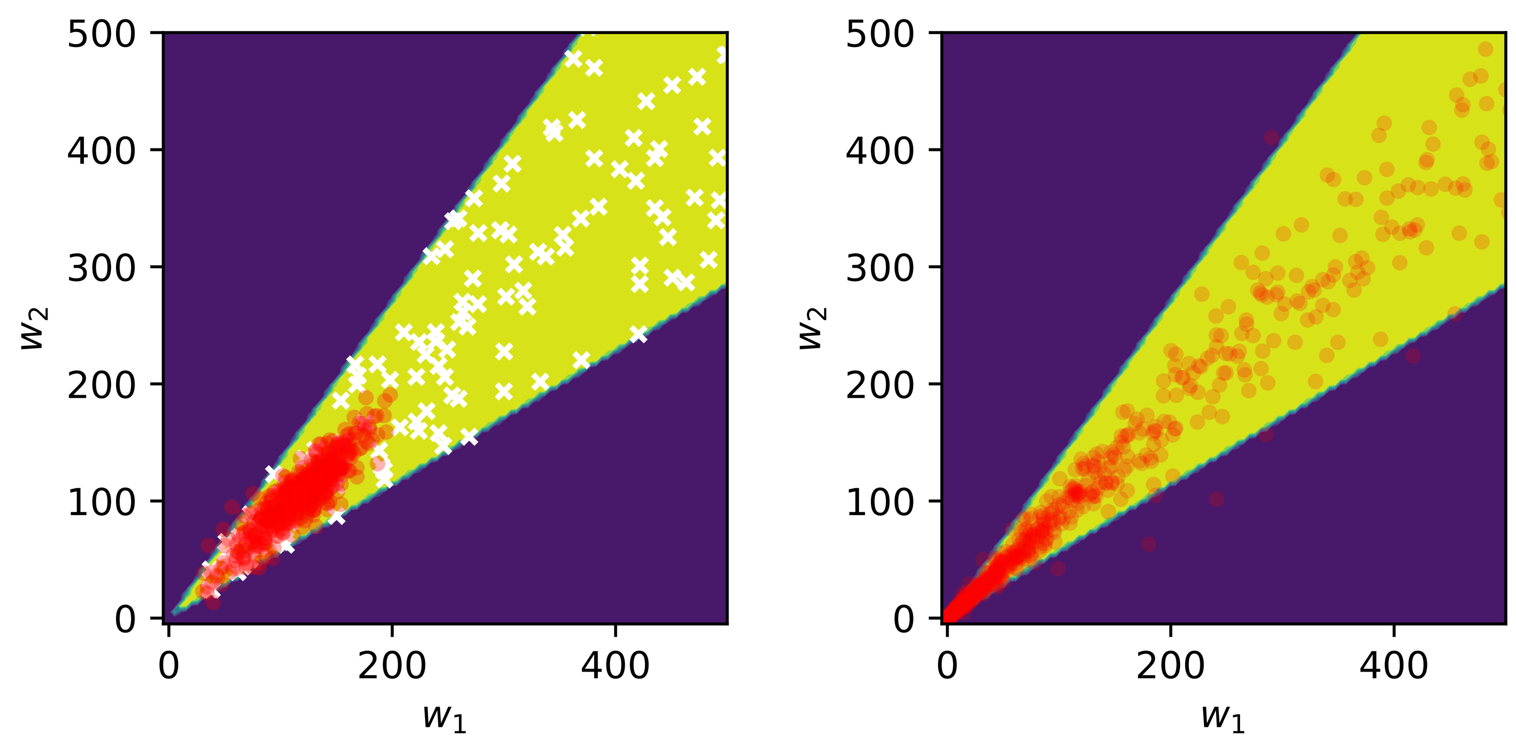
Model	Confidence ↑	Adv. err. ↓	Resampled err. ↓
MAP	1.000	0.999	–
Laplace	0.501	0.499	–
Bootstrap	1.000	0.961	0.957
MCMC	0.976	0.558	0.205
SVI (MC)	0.991	0.606	0.516
Hier. SVI (MC)	0.978	0.678	0.561
MCMC ($\mu \neq 0$)	1.000	0.341	0.301

- **Confidence:** average prob. assigned to the *correct* label on val. set
- **Adv. error:** prob. assigned to the *wrong* label *in the worst case*
- **Resampled error:** adv. error of a *new* ensemble on the same points

Discussion

1. Adv. examples present in a linear model. Regularization not helpful
2. **Accurate Bayesian method (MCMC) makes the model uncertain for adversarial examples**, while remaining confident on validation samples
3. **Bootstrap uncertainty is insufficient in this setup**
4. MCMC results are improved by using a hierarchical prior that exploits symmetry in the data
5. Cheaper, less accurate Bayesian method (SVI) is sufficient for detecting adversarial examples in this setting

Variational posterior



- Variational posterior results in a good predictive distribution, **while not matching the true posterior very well**
- Can use a hierarchical model to try to improve the fit, **but it doesn't necessarily lead to a better predictive distribution**