Adversarial Approaches to Bayesian Learning and Bayesian Approaches to Adversarial Robustness

Ian Goodfellow, OpenAI Research Scientist

NIPS 2016 Workshop on Bayesian Deep Learning

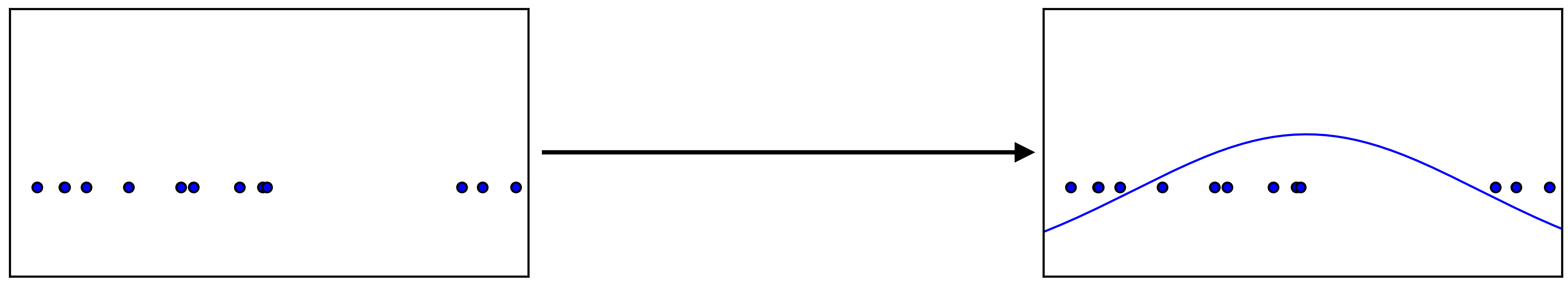
Barcelona, 2016-12-10

Speculation on Three Topics

- Can we build a generative adversarial model of the posterior over parameters?
- Adversarial variants of variational Bayes
- Can Bayesian modeling solve adversarial examples?

Generative Modeling

- Density estimation

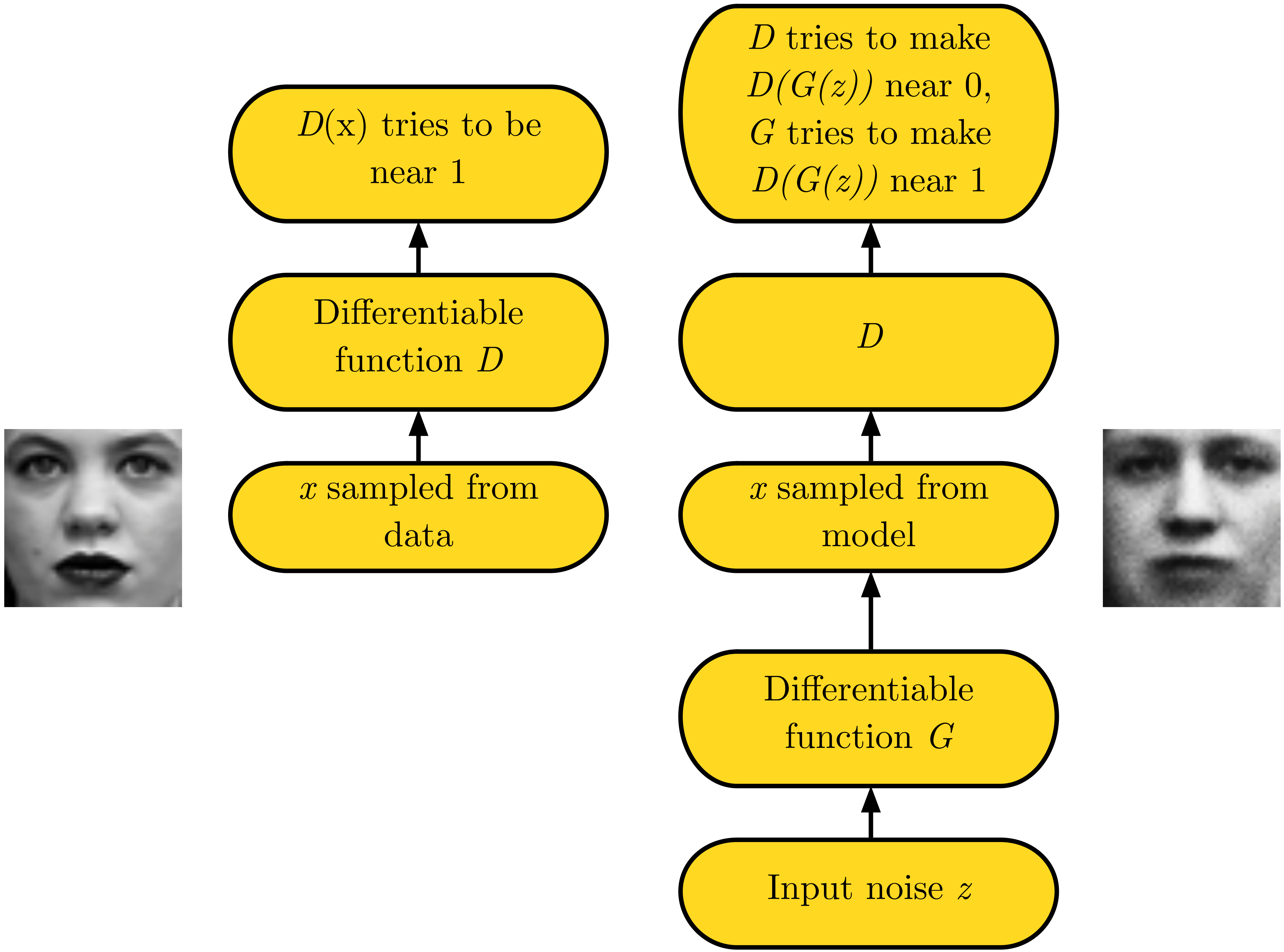


- Sample generation

Training examples Model samples

(Goodfellow 2016)

Adversarial Nets Framework



(Goodfellow 2016)

Minimax Game

$$J^{(D)} = -\frac{1}{2}\mathbb{E}_{\boldsymbol{x}\sim p_{\text{data}}} \log D(\boldsymbol{x}) - \frac{1}{2}\mathbb{E}_{\boldsymbol{z}} \log\left(1 - D\left(G(\boldsymbol{z})\right)\right)$$
$$J^{(G)} = -J^{(D)}$$

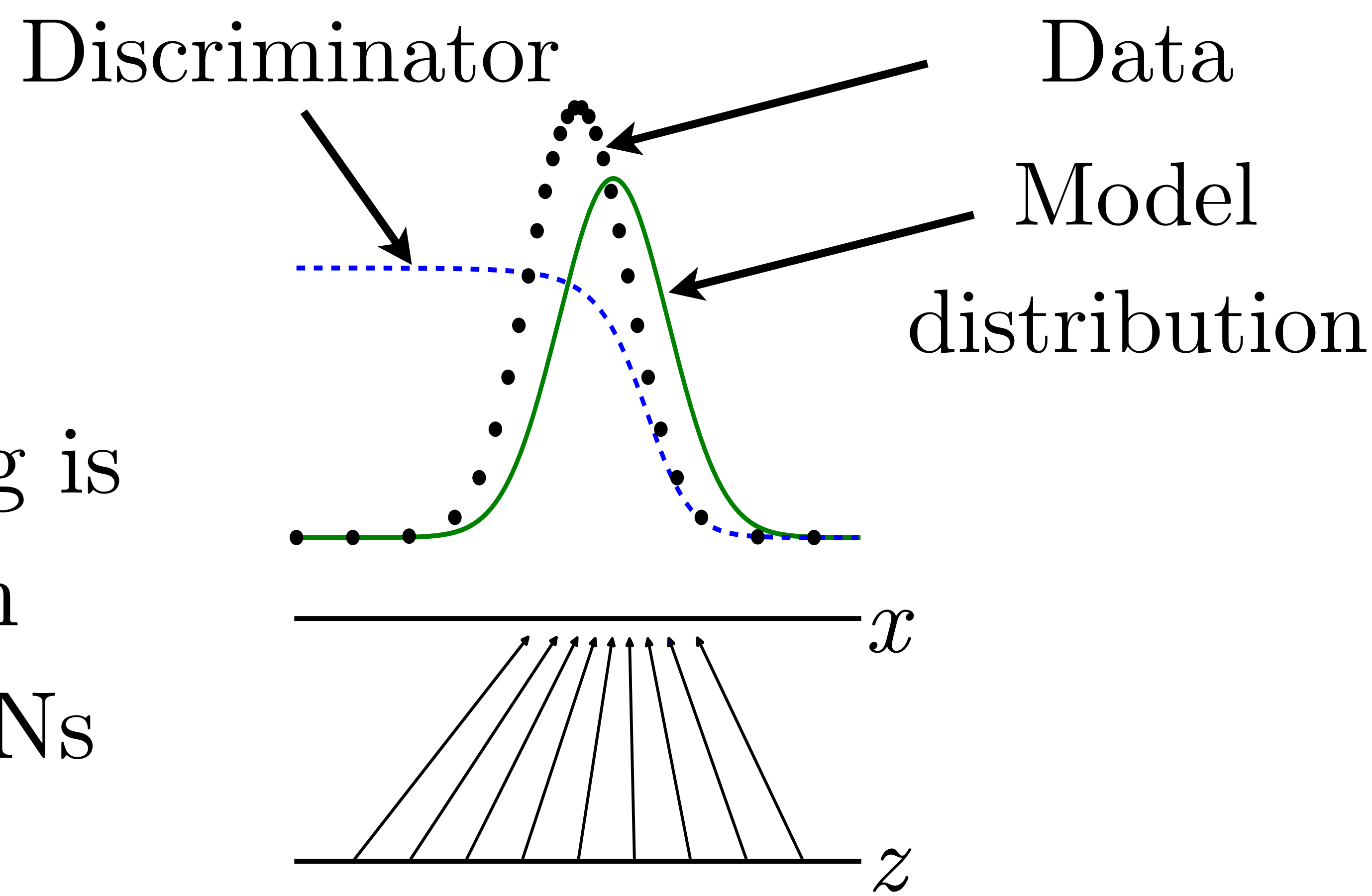
-Equilibrium is a saddle point of the discriminator loss
-Resembles Jensen-Shannon divergence
-Generator minimizes the log-probability of the discriminator being correct

(Goodfellow 2016)

Discriminator Strategy

Optimal $D(\boldsymbol{x})$ for any $p_{\text{data}}(\boldsymbol{x})$ and $p_{\text{model}}(\boldsymbol{x})$ is always

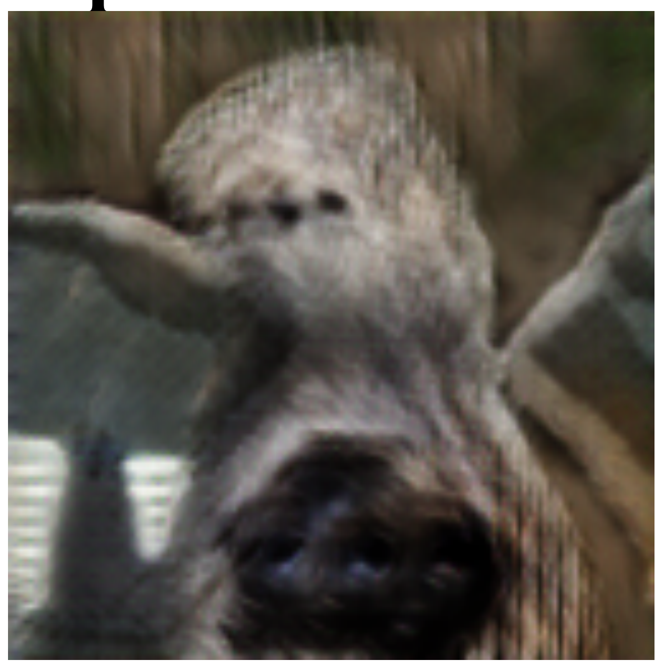
$$D(x) = \frac{p_{\text{data}}(x)}{p_{\text{data}}(x) + p_{\text{model}}(x)}$$

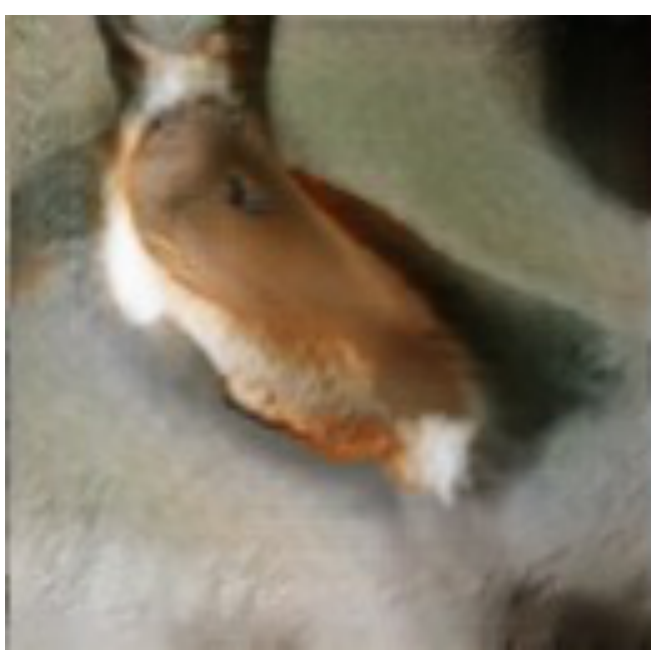


Estimating this ratio using supervised learning is the key approximation mechanism used by GANs

(Goodfellow 2016)

High quality samples from complicated distributions

(Goodfellow 2016)

Speculative idea: generator nets for sampling from the posterior

- Practical obstacle:
  - Parameters lie in a much higher dimensional space than observed inputs
- Possible solution:
  - Maybe the posterior does not need to be extremely complicated
  - HyperNetworks (Ha et al 2016) seem to be able to model a distribution on parameters

Theoretical problems

- A naive application of GANs to generating parameters would require samples of the parameters from the true posterior

- We only have samples of the data that were generated using the true posterior

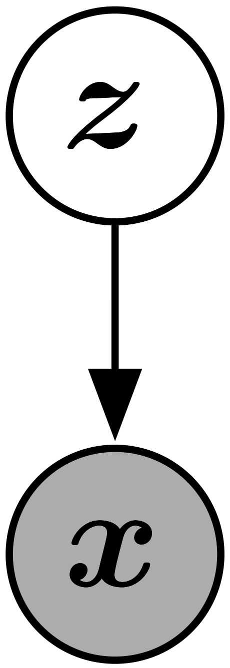
(Goodfellow 2016)

HMC approach?

$$\frac{p(\boldsymbol{X} \mid \boldsymbol{\theta})}{p(\boldsymbol{X} \mid \boldsymbol{\theta}^*)} = \Pi_i \frac{p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta})}{p(\boldsymbol{x}^{(i)} \mid \boldsymbol{\theta}^*)}$$

- Allows estimation of unnormalized likelihoods via discriminator
- Drawbacks:
  - Discriminator needs to be re-optimized after visiting each new parameter value
  - For the likelihood estimate to be a function of the parameters, we must include the discriminator learning process in the graph for the estimate, as in unrolled GANs (Metz et al 2016)

Variational Bayes

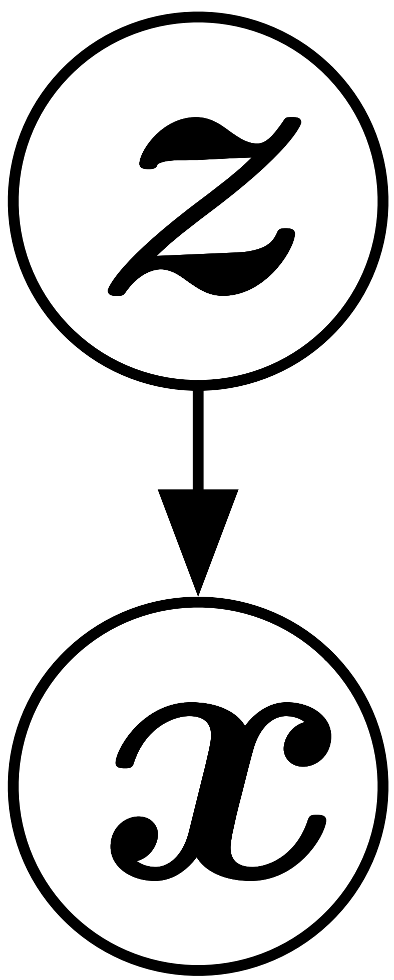


$$
\begin{aligned}
\log p(\boldsymbol{x}) \geq & \log p(\boldsymbol{x}) - D_{\mathrm{KL}}\left(q(\boldsymbol{z}) \| p(\boldsymbol{z} \mid \boldsymbol{x})\right) \\
= & \mathbb{E}_{\boldsymbol{z} \sim q} \log p(\boldsymbol{x}, \boldsymbol{z}) + H(q)
\end{aligned}
$$

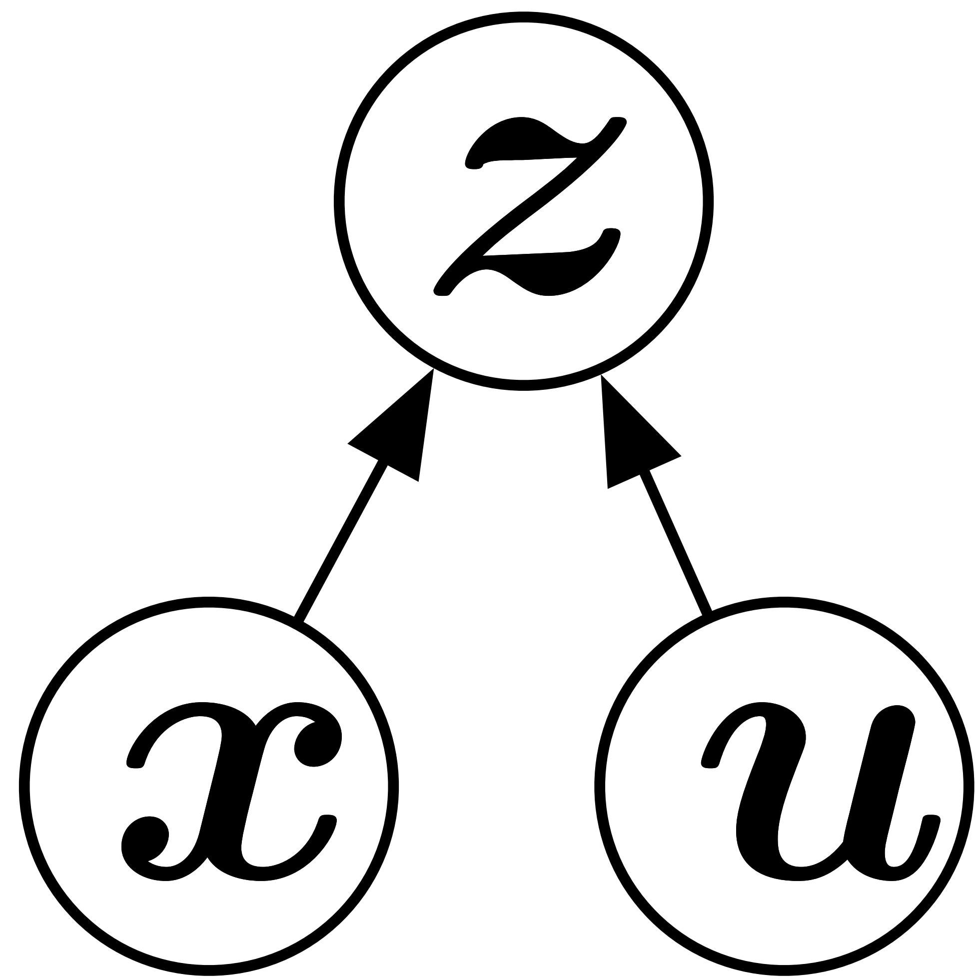
- Same graphical model structure as GANs
- Often limited by expressivity of q

(Goodfellow 2016)

Arbitrary capacity posterior via backwards GAN



Generation process



Posterior sampling process

(Goodfellow 2016)

Related variants

- Adversarial autoencoder (Makhzani et al 2015)
  - Variational lower bound for training decoder
  - Adversarial training of encoder
    - Restricted encoder
    - Makes aggregate approximate posterior indistinguishable from prior, rather than approximate posterior indistinguishable from true posterior
    - Uses variational lower bound for training decoder

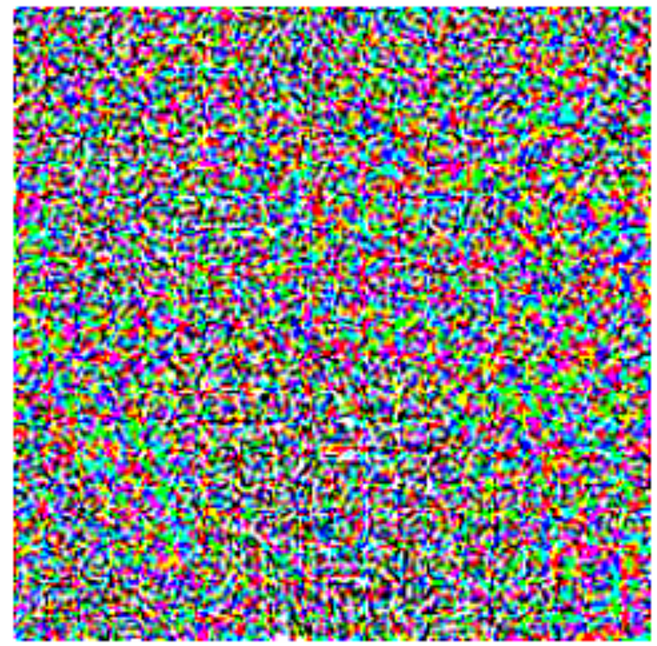
ALI / BiGAN

- Adversarially Learned Inference (Dumoulin et al 2016)
  - Gaussian encoder
- BiGAN (Donahue et al 2016)
  - Deterministic encoder

Adversarial Examples

panda
58% confidence

$+ .007 \times$

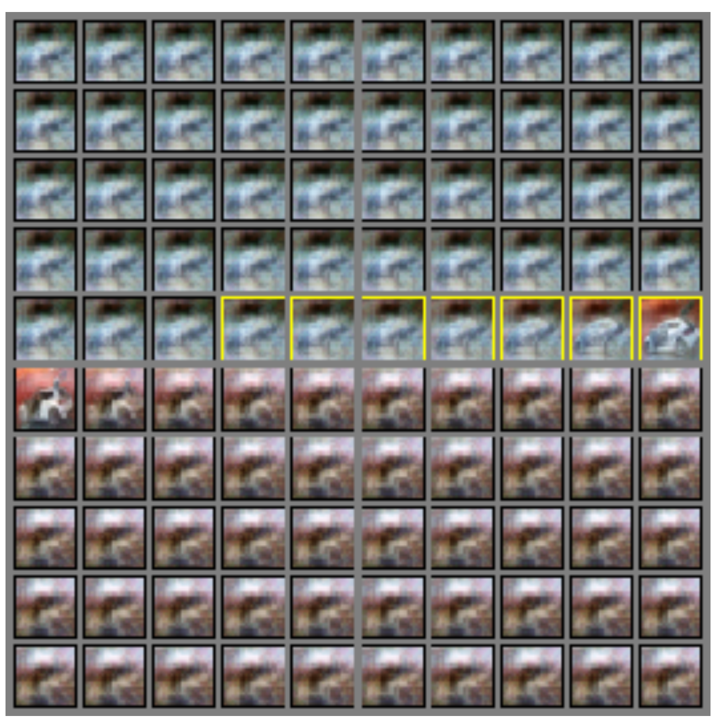


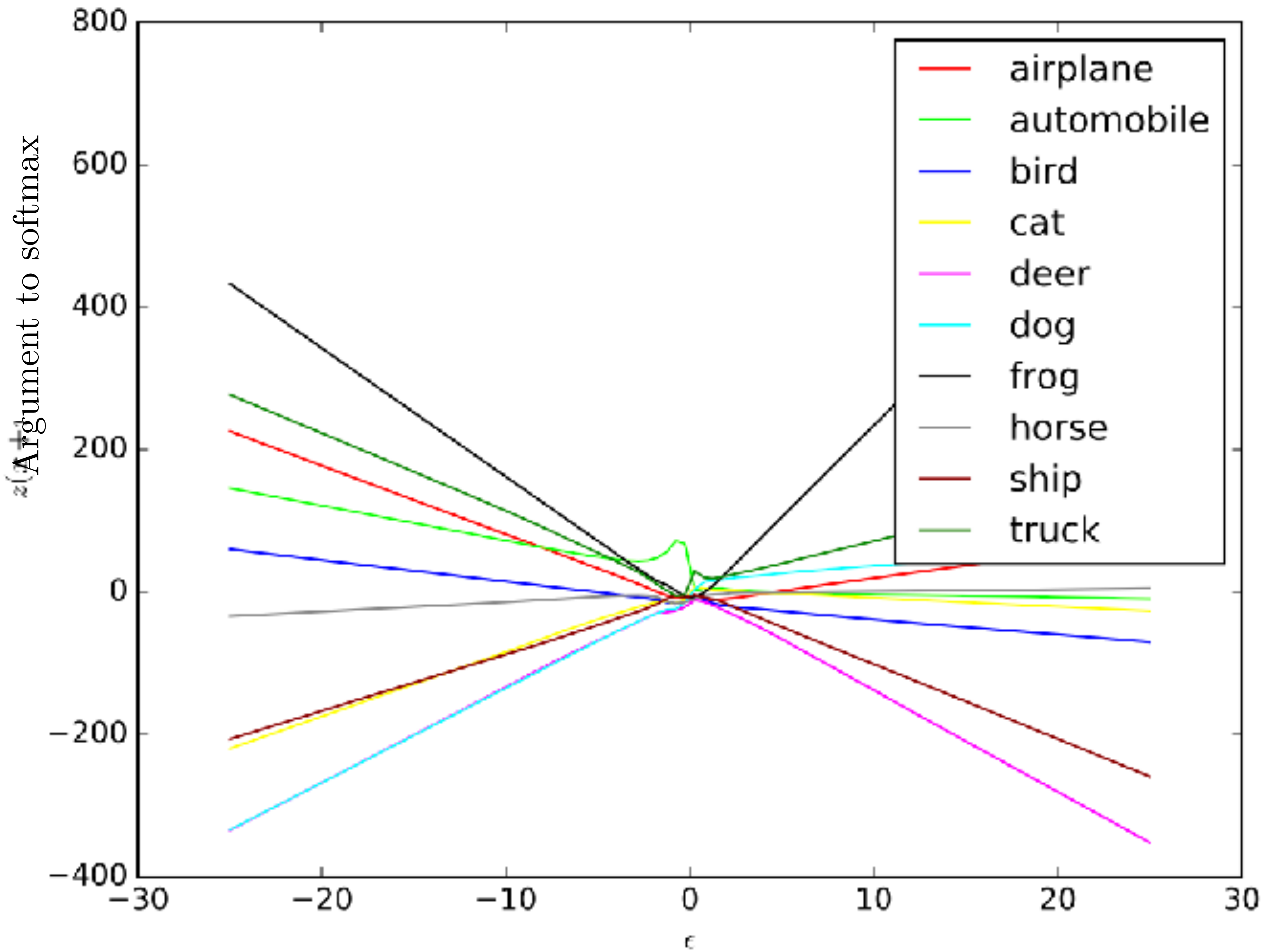
$=$

gibbon
99% confidence

(Goodfellow 2016)

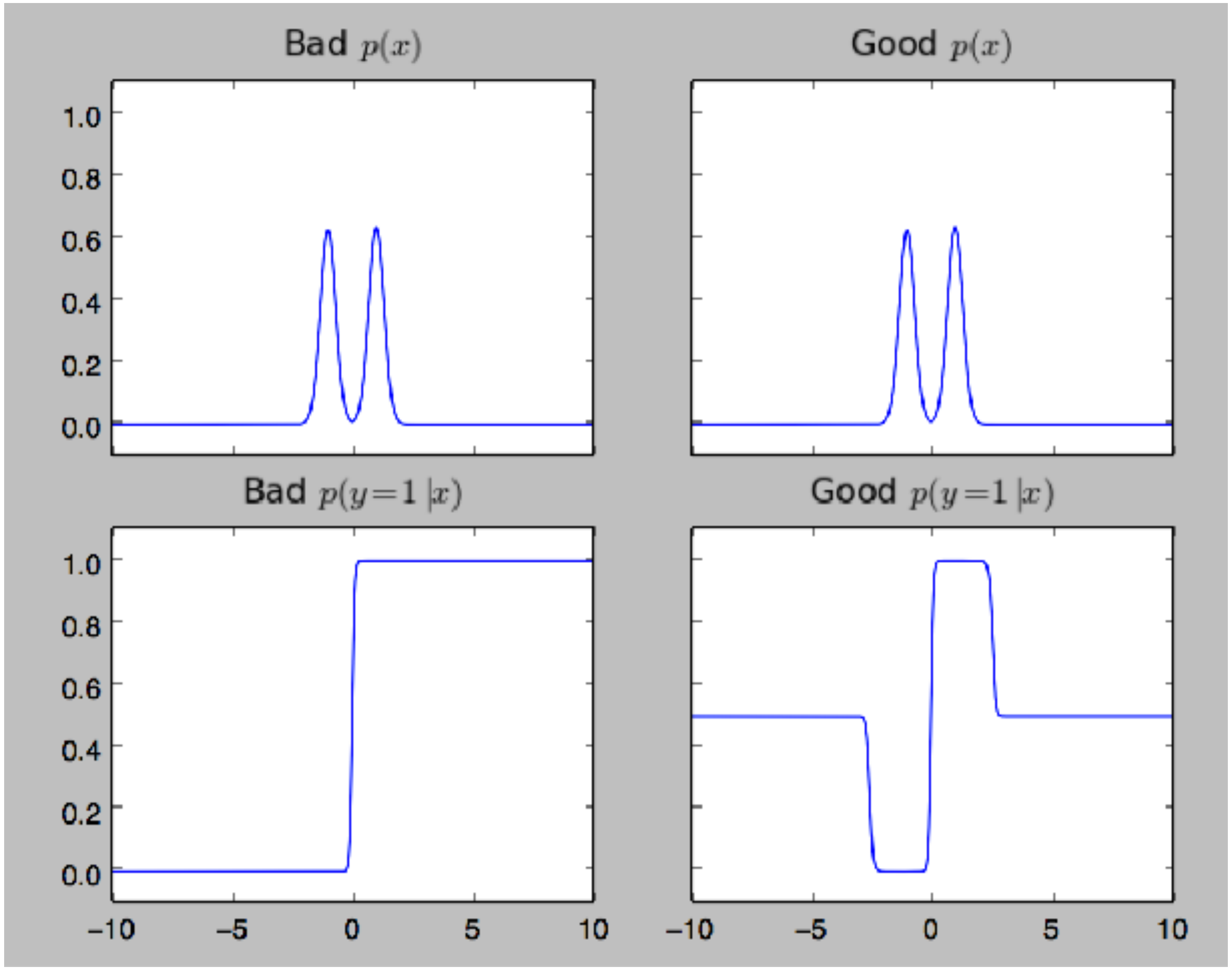
Overly linear, increasingly confident extrapolation



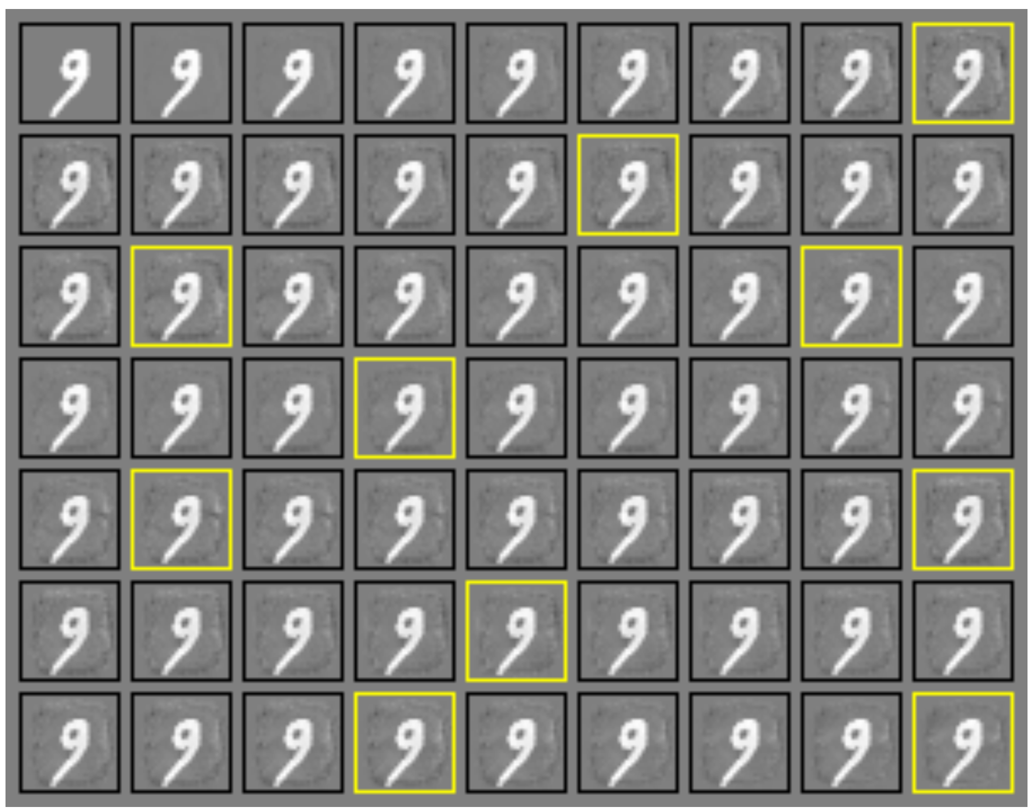


Designing priors on latent factors

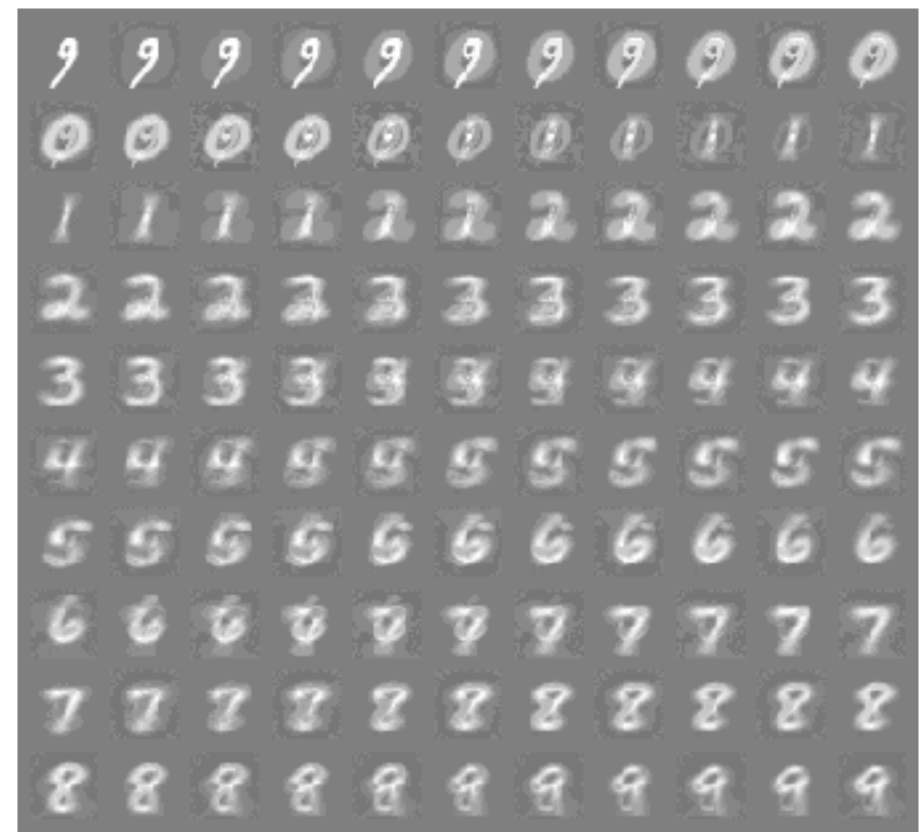
Both these two class mixture models implement roughly the same marginal over x, with very different posteriors over the classes. The likelihood criterion cannot strongly prefer one to the other, and in many cases will prefer the bad one.



RBFs are better than linear models



Attacking a linear model



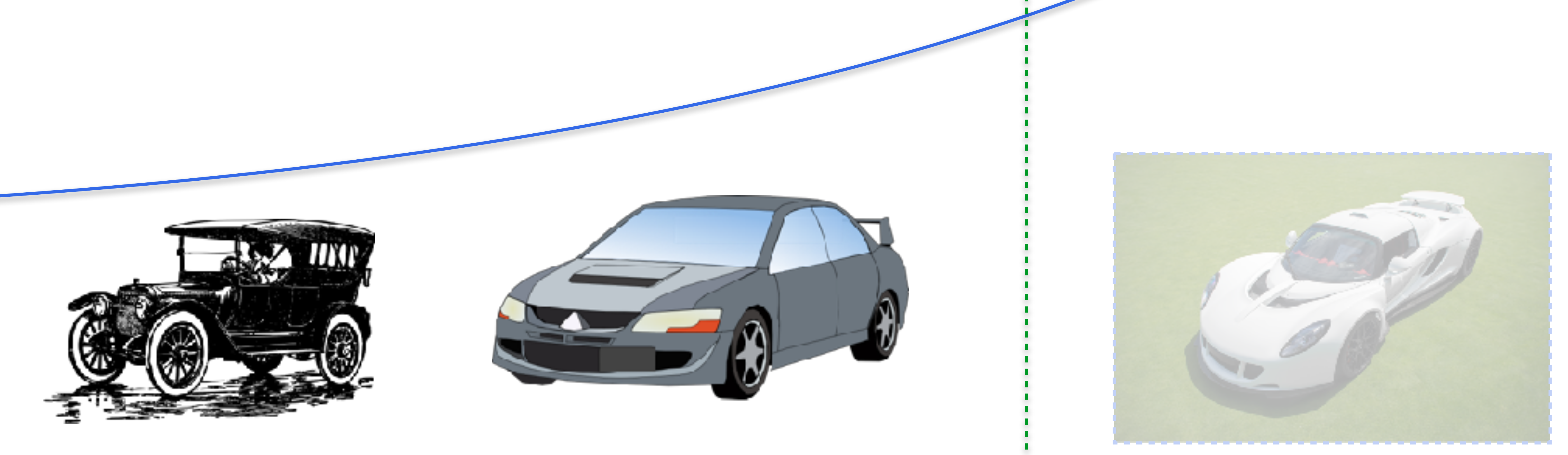
Attacking an RBF model

Possible Bayesian solutions

- Bayesian neural network
  - Better confidence estimates might solve the problem
  - So far, has not worked, but may just need more effort
    - Variational approach
    - MC dropout
- Regularize neural network to emulate Bayesian model with RBF kernel (amortized inference of Bayesian model)

Universal engineering machine (model-based optimization)

Make new inventions by finding input that maximizes model's predicted performance



(Goodfellow 2016)

Conclusion

- Generative adversarial nets may be able to
    - Sample from the Bayesian posterior over parameters
    - Implement an arbitrary capacity q for variational Bayes
- Bayesian learning may be able to solve the adversarial example problem and unlock the potential of model-based optimization