ROBUSTNESS OF BAYESIAN APPROACH TO GRADIENT-BASED ATTACKS

Objective: *Demonstrate the theoretical robustness of Bayesian neural architectures against multiple white-box attacks*

Sowmya Jayaram Iyer

jayarami@purdue.edu

Bayesian Neural Networks

- Training a neural network via optimization is (from a probabilistic perspective) equivalent to maximum likelihood estimation (MLE) for the weights.
- This ignores any uncertainty that we may have in the proper weight values.
- As NNs often do, this may be a reason for overfitting.
- Partial fix: Regularization a.k.a inducing priors on the weights from a Bayesian perspective.
- A theoretically justifiable way for people who love statistics and probabilities would be to go with posterior inference. However, intractable.
- There have been some interesting recent results using **Variational Inference** to do this.

Variational Inference

- **“Inference”:** We want to infer the latent variables (W) from observed data (X).
- In Bayesian modeling, we want to be able to sample from the posterior of models given the data.
- How do we obtain P(W|X=D

  By Bayes’ Theorem,

$$P(W|X = D) = \frac{\overbrace{P(X = D|W)}^{\text{Likelihood}} \cdot \overbrace{P(W)}^{\text{Prior}}}{\underbrace{P(X = D)}_{\text{Marginal}}}$$

- This marginal is computationally intractable.

$$P(X) = \int_{W_0} \dots \int_{W_D} P(X, W) dW_0 \dots dW_D$$

Variational Inference and ELBO

- Instead we find a simple distribution $q(W) \approx p(W|X = D)$
- That's a **Variational** problem as we want to optimize it for a function to be able to perform Inference.
- When we say optimize, we need a Loss function to find how close this simple distribution is to the posterior distribution. (KL-Divergence)
- Our task is now to minimize this distance, that is, KL-Divergence.

$$q^*(W) = argmin_{q(W)\in Q} KL(q(W)||p(W|D))$$

- But, we don't have the posterior. Rearranging the above, we get:

ELBO formulation

$$\underbrace{KL(q(W)\ ||\ p(W|D))}_{\text{Distance Metric : +ve}} = \int_W q(W)\ log\left(\frac{q(w)\cdot p(D)}{p(W,D)}\right)dW$$

$$= \int_W q(W)\cdot log\left(\frac{q(w)}{p(W,D)}\right)dW + \int_W q(W)\ log(p(D))dW$$

$$= \underbrace{-E_{w\sim q(w)} log\left(\frac{p(W,D)}{q(w)}\right)}_{\text{Evidence Lower Bound}} + \underbrace{log(p(D))}_{\text{Evidence : Fixed quantity and something negative}}$$

$$q^*(W) = argmax_{q(W)\in Q} ELBO)$$

Hence, our Objective function becomes ELBO:

$$\mathcal{F}(\mathcal{D},\theta) \approx \sum_{i=1}^{n} \log q_\theta(w^{(i)}|\mathcal{D}) - \log p(w^{(i)}) - \log p(\mathcal{D}|w^{(i)})$$

The variational posterior is taken as Gaussian distribution centered around mean μ and variance as σ^2.

$$log\ q(W/D) = \sum_i log\ \mathcal{N}(W|\mu,\ \sigma^2)$$

The prior is taken as individual Gaussians. $log\ p(W) = \sum_i log\ \mathcal{N}(W|0,\ \sigma_p^2)$

Experiment

- MODELS:
  - Frequentist: AlexNet, LeNet and a simple CNN
  - Bayesian: BAlexNet, BLeNet, BCNN
- DATASETS:
  - CIFAR10, MNIST
- Adversarial Attacks:
  - FGSM (Fast Sign Gradient Method)
  - PGD (Projected Gradient Descent)
  - BIM (Basic iterative method or Iterative-FSGM)

Gradient- Based Adversarial Attack

- Given an input point x* and a strength (i.e. maximum perturbation magnitude) > 0, the worst-case adversarial perturbation can be defined as the point around x* that maximizes the loss function.

$$\bar{x} = argmax_{\bar{x}:||\bar{x}-x*||<\epsilon} L(\bar{x}, w)$$

- If network prediction on $\bar{x}$ differs from the original label, this implies that $\bar{x}$ is an adversarial example and the attack was successful.
- In particular, the FGSM is among the most commonly employed Gradient-Based attacks. In the context of BNNs, where attacks are against the predictive distribution

$$\tilde{\mathbf{x}} \simeq \mathbf{x} + \epsilon \,\mathrm{sgn}\left(\langle \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)}\right) \simeq \mathbf{x} + \epsilon \,\mathrm{sgn}\left(\sum_{i=1}^{n} \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}_i)\right)$$

Reason for BNN robustness

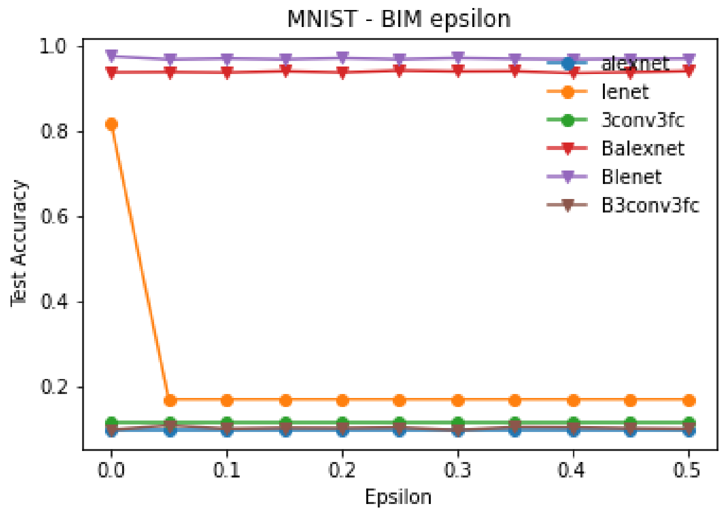
$$\tilde{\mathbf{x}} \simeq \mathbf{x} + \epsilon \,\mathrm{sgn}\left(\langle \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)}\right) \simeq \mathbf{x} + \epsilon \,\mathrm{sgn}\left(\sum_{i=1}^{n} \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}_i)\right)$$

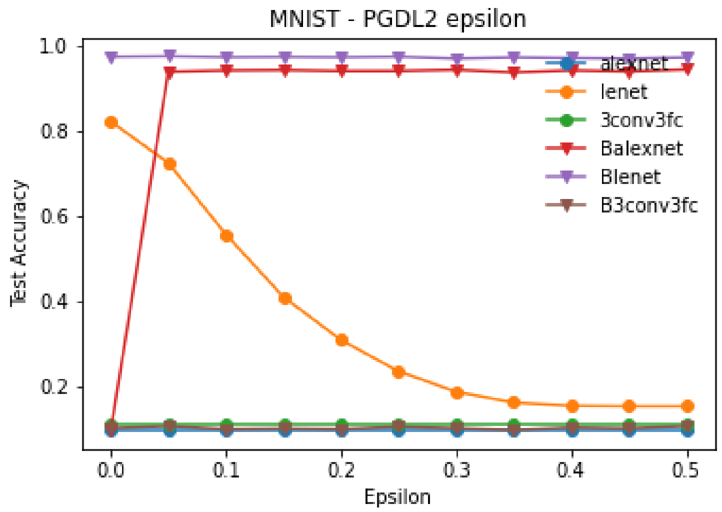
- For any such gradient attacks on BNN, the samples w_i are drawn from the posterior probability.
- A possible explanation for robustness is that the above averaging under the posterior might lead to cancellations in the final expectation of the gradients.
- [Rotskoff and Vanden-Eijnden, 2018] proved global convergence of (stochastic) gradient descent (at the distributional level) in the over parametrized, large data limit.
- By the definition of the FGSM attack and other gradient-based attacks, Theorem 1 directly implies that any gradient-based attack will be ineffective against a BNN in the large data limit.

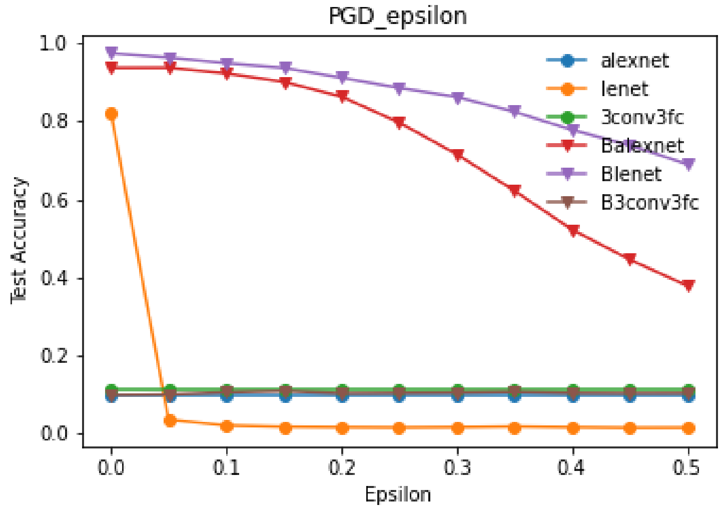
Theorem 1. *Let $f(\mathbf{x}, \mathbf{w})$ be a fully trained overparametrized BNN on a prediction problem with data manifold $\mathcal{M}_D \subset \mathbb{R}^d$ and posterior weight distribution $p(\mathbf{w}|D)$. Assuming $\mathcal{M}_D \in C^\infty$ almost everywhere, in the large data limit we have a.e. on $\mathcal{M}_D$*

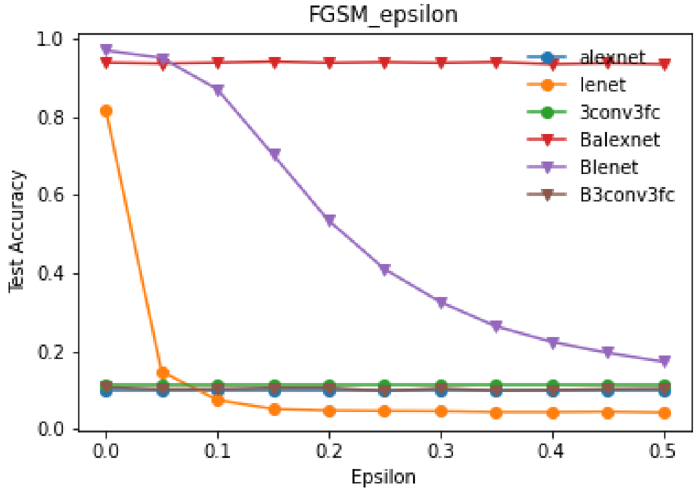
$$\left(\langle \nabla_{\mathbf{x}} L(\mathbf{x}, \mathbf{w}) \rangle_{p(\mathbf{w}|D)}\right) = \mathbf{0}. \quad (3)$$

Results on MNIST - AlexNet and LeNet









Results on CIFAR10