1 Robustness

1. Suppose $x_{\text{adv}} \in \mathbb{R}^d$ and $(x_{\text{adv}})_i = x_i + \varepsilon$. What is $\|x - x_{\text{adv}}\|_p$ for $p = 1, 2, \infty$? Is the perturbation norm larger for $p = 1$ or $p = \infty$?

2. Mixup, AugMix, and PixMix use Beta or Dirichlet distributions. These distributions require hyperparameters. Given the Beta distributions $\text{Beta}(1, 1)$, $\text{Beta}(5, 5)$, $\text{Beta}(0.5, 0.5)$, which distribution is has a peak at its center, has a peak at its corners, and is a uniform distribution?

3. Suppose $\mathbf{x}$ is our original input, y is its original label, θ is our model parameters, and $\mathcal{L}$ is the cross entropy loss. Let $l(\mathbf{x})_i$ be the logit (pre-softmax input) from our model at dimension i, given input $\mathbf{x}$.

   Consider a "MultiTargeted" attack which targets each wrong class out of K possible classes:

   ```
   For $t \in \{1, \dots, y-1, y+1, \dots, K\}$
       $\mathcal{L}(\mathbf{x}, y; \theta) := l(\mathbf{x})_y - l(\mathbf{x})_t$
       $\mathbf{x}_{\text{adv,t}} = x + \mathcal{O}(t)$
   $\mathbf{x}_{\text{adv}}$ is set to be whichever $\mathbf{x}_{\text{adv,t}}$ incurs the greatest loss.
   ```

   Mark true for all of the following statements that are correct about this attack, and false otherwise.

   (a) We should set $\mathcal{O}(t) = \text{argmin}_{\delta:\|\delta\|_p \le \varepsilon} \mathcal{L}(\mathbf{x} + \delta, y; \theta)$.

   (b) We should set $\mathcal{O}(t) = \text{argmax}_{\delta:\|\delta\|_p \le \varepsilon} \mathcal{L}(\mathbf{x} + \delta, y; \theta)$.

   (c) This attack can be at least $K - 1$ times as expensive as an untargeted attack.

   (d) The MultiTargeted attack has some of the drawbacks of untargeted attacks, e.g., it can cause models to misclassify examples as similar classes.

4. The minimum ℓ_2 norm attack x_{adv} for a two class classifier is given by $\min_x \|x - x_0\|_2$ subject to $w^\mathsf{T} x + w_0 = 0$. (Assume $w \neq 0$.)

   Which is true?

   (a) $x_{\text{adv}} = 0$

   (b) $x_{\text{adv}} = x_0 - (w^\mathsf{T} x_0 + w_0)\frac{w}{\|w\|_2}$

   (c) $x_{\text{adv}} = x_0 - \left(\frac{w^\mathsf{T} x_0 + w_0}{\|w\|_2}\right)\frac{w}{\|w\|_2}$

(d) $x_{\text{adv}} = x_0 - (w^\mathsf{T} x_0 + w_0)w$

5. Suppose we use adversarial training to train our network. The adversarial examples are generated through the fast gradient sign method (FGSM) and are within an ε distance (according to the ℓ_∞ norm) from our original points. Mark true for all of the following statements that are correct about the resulting model, and false otherwise.

   (a) The network is certified to be robust against all adversarial examples that are less than an ε distance (according to the ℓ_∞ norm) from the original training points.

   (b) Compared to a classifier trained without adversarial training, our classifier will likely have a higher accuracy on non-adversarial examples.

   (c) Compared to a classifier trained without adversarial training, our classifier will likely have a higher accuracy on adversarial examples generated via FGSM.

   (d) The classifier is likely to be more robust to adversarial examples generated from FGSM compared to adversarial examples generated through a different method, e.g. PGD.

6. Mark true for all of the following statements that are correct about adversarial examples and adversarial defense, and false otherwise.

   (a) Suppose we have set up a normal training process and trained a model. In order to defend against adversarial examples, it suffices to first generate a set of adversarial examples from our already trained model, add them to the dataset we have, and then train a new model from scratch using the same training process we already have.

   (b) When generating adversarial examples to attack a model, if we cannot have access to that model directly, we can train a similar model with the same dataset and generate adversarial examples on this new model. The adversarial examples can sometimes be effective against the original model.

   (c) Suppose we want to generate an untargeted adversarial example for a given model by maximizing the loss and we want to constrain the norm of our perturbation to be less than ε. If we constrain the perturbation's ℓ_∞ norm to be less than ε, we can typically generate an stronger adversarial example that induces greater loss compared to the adversarial example we can generate if we constraint the perturbation's ℓ_2 norm to be less than ε.

   (d) Suppose we first train our model and then make our model non-differentiable by finely quantizing the activations for each layer without changing the model's prediction. Now since we cannot take gradients with respect to the model input, our model is guaranteed to be safe against any adversarial examples.

7. Which of the following are true?

(a) The only difference between FGSM and PGD is that PGD uses multiple steps.

(b) Adversaries trying to upload copyrighted content to YouTube do not always need to constrain their content modifications to a small ℓ_p perturbation.

(c) Suppose we are training a logistic regression model with $y \in \{-1, 1\}$, $\min_{\mathbf{w}} \mathbb{E}_{(\mathbf{x},y)\sim p_{\text{data}}} \mathcal{L}(\mathbf{x}, y; \mathbf{w})$, where $\mathcal{L}(\mathbf{x}, y; \mathbf{w}) = \log(1 + e^{-y\mathbf{w}^\mathsf{T}\mathbf{x}})$.
Then $\text{sign}(\nabla_x \mathcal{L}(\mathbf{x}, y; w)) = -y\text{sign}(\mathbf{w})$.

(d) Targeted attacks are usually better able to reduce accuracy than untargeted attacks.

(e) White box attacks are usually better able to reduce accuracy than black box attacks

8. Mark true for all of the following statements that are correct.

(a) ImageNet challenge datasets (e.g., ImageNet-C,R,A) contain image classes that are different from the ones in ImageNet. Thus, they are a good stress test of whether models trained on ImageNet can also generalize to different classes.

(b) The ANLI dataset consists of adversarial examples for natural language inference that were generated through a GAN.

(c) The ImageNet-C paper shows that models can't be made more robust to Gaussian noise by training them with training data augmented with Gaussian noise.

(d) Mixup is used for augmenting language modeling text datasets

(e) PixMix mixes images (possibly fractals) and augmented images together