Assignment 1 — White-Box Attack (FGSM)

Submission Instructions

This assignment is due **Sunday, September 21, 2025, by 11:59 pm**. Please submit your solutions via Canvas. You should submit a PDF of your write-up along with the complete source code. Please do not include blurry scanned/photographed equations, as they are difficult for us to grade.

Late Submission Policy

The late submission policy for assignments will be as follows unless otherwise specified:

1. 75% credit within 0-48 hours after the submission deadline.

2. 50% credit within 48-96 hours after the submission deadline.

3. 0% credit 96 hours after the submission deadline

Overview

Implement and evaluate **Fast Gradient Sign Method (FGSM)** attacks under an L∞ threat model on two architectures (ResNet-18, ViT) and two datasets (MNIST, CIFAR-10). Compare targeted vs. untargeted objectives and report robustness metrics with clear, reproducible experiments.

Learning goals

- Understand L∞ and **targeted vs. untargeted** objectives
- Implement FGSM correctly
- Evaluate robustness across architectures (ResNet-18 vs. ViT) and datasets (MNIST vs. CIFAR-10)

Datasets & models

- **Datasets:** MNIST (28×28), CIFAR-10 (32×32)
- **Models (both datasets): ResNet-18** and a **ViT** (tiny/small)

Tips

- ViT on small images: use a small ViT configured for 32×32 (CIFAR-10) and 28×28 (MNIST), or upsample to 224 and use ViT-Tiny/16. Keep parameters modest.

Tasks

1. **Train or load baseline models**
    - MNIST: ≥ **95%** clean accuracy

    - CIFAR-10: ResNet-18 ≈ **85%+**, ViT ≈ **80%+**
      (Targets, not hard requirements; report your actual clean accuracy.)
2. **Implement FGSM (L∞) from scratch**
3. **Calibrate ε (L∞)**
   Evaluate ε ∈ {1/255, 2/255, 4/255, 8/255}.
4. **Evaluate on a fixed test subset** (e.g., **1,000 images**) for each **(dataset × model)**:
    - **Clean Accuracy**
    - **Robust Accuracy** under **untargeted FGSM** for each ε
    - **Attack Success Rate (ASR)**
    - **Targeted ASR** for two target choices: (i) random target, (ii) least-likely class (chosen from clean logits)

Deliverables

1. **Code** (FGSM implementation + evaluation script) and a short **README** with exact commands to reproduce results.
2. **2–4 page write-up** including:
    - **Table (per dataset/model):** Clean Acc; Robust Acc at ε ∈ {1,2,4,8}/255; **ASR (untargeted)**; **Targeted ASR** (random & least-likely)
    - **Figure (required):** for each dataset/model and attack type, show **Original | Perturbation | Adversarial** for selected samples (e.g., 10 examples at ε=8/255). Visualize the perturbation as the **difference**, scaled for visibility (e.g., ×10), with labels/confidences.
    - **Brief analysis (≤1 page):** differences between ResNet-18 and ViT, targeted vs. untargeted behavior, notable failure cases or sensitivities (e.g., normalization, ε).

Rubric (100 pts)

- **Correct FGSM (untargeted & targeted)** — 30
- **Experimental** (metrics complete, ε sweep, fixed subset) — 25
- **Results quality** (clear tables/plots, per-ε trends) — 20
- **Visualizations** (original/perturbation/adversarial with labels/confidence) — 15
- **Clarity & reproducibility** (seeds, versions, README, exact commands) — 10