1 Adversarial Attacks

- Targeted FGSM: $x' = x - \epsilon\,\mathbf{sgn}(\nabla_x \mathcal{L}_t(x))$, where t is the target label. Untargeted FGSM: $x' = x + \epsilon\,\mathbf{sgn}(\nabla_x \mathcal{L}_y(x))$, where y is the original label.
- CW: use L-BFGS to solve $\mathbf{argmin}_\eta \|\eta\|_p + c \cdot \mathrm{obj}_t(x+\eta)$, s.t. $x + \eta \in [0,1]^n$.