ECE595 / STAT598: Machine Learning I
Lecture 35 Max-Loss Attacks and Regularized Attacks

Spring 2020

Stanley Chan

School of Electrical and Computer Engineering
Purdue University

PURDUE UNIVERSITY

Agenda

- Last lecture we have seen min-distance attack
- In linear case, there is a very simple geometry
- Today we are going to consider two of its variations
  - Max-loss attack
  - Regularized attack
- We will again talk about their geometry using **linear** models.
- And then we will link the results to deep models.
- You will see that some of the most popular deep attack models out there are based on one of the three formulations we discuss here

Outline

- Lecture 33 Overview
- Lecture 34 Min-distance attack
- Lecture 35 Max-loss attack and regularized attack

Today's Lecture

- Max-loss attack
  - Linear models
  - Deep models: FGSM and PGD
- Regularized attack
  - Linear models
  - CW attack

Maximum Loss Attack

Definition (Maximum Loss Attack)

The **maximum loss attack** finds a perturbed data $\boldsymbol{x}$ by solving the optimization

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{maximize}} \quad & g_t(\boldsymbol{x}) - \max_{j \neq t} \{g_j(\boldsymbol{x})\} \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta, \end{aligned} \tag{1}$$

where $\|\cdot\|$ can be any norm specified by the user, and $\eta > 0$ denotes the attack strength.

- I want to bound my attack $\|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta$
- I want to make $g_t(\boldsymbol{x})$ as big as possible
- So I want to maximize $g_t(\boldsymbol{x}) - \max_{j \neq t} \{g_j(\boldsymbol{x})\}$
- This is equivalent to

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta, \end{aligned}$$

If you restrict yourself to two classes only ...

- The problem is

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta, \end{aligned}$$

- η is the maximum loss attack strength
- Want $g_t(\boldsymbol{x})$ to override $\max_{j \neq t} \{g_j(\boldsymbol{x})\}$
- So maximize $g_t(\boldsymbol{x})$
- If you restrict to linear, and only two classes, then

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{w}^T \boldsymbol{x} + w_0 \quad \text{subject to} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta.$$

- Solvable in closed-form.

Max-Loss Attack using ℓ_2-norm

- The problem is

$$\underset{\boldsymbol{r}}{\text{minimize}} \quad \boldsymbol{w}^T\boldsymbol{r} + b_0 \quad \text{subject to} \quad \|\boldsymbol{r}\|_2 \le \eta.$$

- Cauchy inequality:

$$\boldsymbol{w}^T\boldsymbol{r} \ge -\|\boldsymbol{w}\|_2\|\boldsymbol{r}\|_2 \ge -\eta\|\boldsymbol{w}\|_2.$$

- Claim: Lower bound of $\boldsymbol{w}^T\boldsymbol{r}$ is attained when $\boldsymbol{r} = -\eta\boldsymbol{w}/\|\boldsymbol{w}\|_2$:

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{r} &= \boldsymbol{w}^T\left(-\frac{\eta\boldsymbol{w}}{\|\boldsymbol{w}\|_2}\right)\\ &= -\eta\|\boldsymbol{w}\|_2.\end{aligned}$$

- So the solution is $\boldsymbol{r} = -\eta\boldsymbol{w}/\|\boldsymbol{w}\|_2$.

Max-Loss Attack using ℓ_∞-norm

- Goal: Want to solve

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{w}^T\boldsymbol{x} + w_0 \quad \text{subject to} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta.$$

- Define $\boldsymbol{x} = \boldsymbol{x}_0 + \boldsymbol{r}$. Then

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{x} + w_0 &= \boldsymbol{w}^T(\boldsymbol{x}_0 + \boldsymbol{r}) + w_0 \\ &= \boldsymbol{w}^T\boldsymbol{x}_0 + \boldsymbol{w}^T\boldsymbol{r} + w_0 \\ &= \boldsymbol{w}^T\boldsymbol{r} + \underbrace{\boldsymbol{w}^T\boldsymbol{x}_0 + w_0}_{=b_0}\end{aligned}$$

- Define $b_0 = (\boldsymbol{w}^T\boldsymbol{x}_0 + w_0)$. The optimization can be rewritten as

$$\underset{\boldsymbol{r}}{\text{minimize}} \quad \boldsymbol{w}^T\boldsymbol{r} + b_0 \quad \text{subject to} \quad \|\boldsymbol{r}\|_\infty \leq \eta.$$

Solution to Max-Loss Attack (ℓ_∞-norm)

- Holder's inequality (the negative side):

$$\boldsymbol{w}^T\boldsymbol{r} \geq -\|\boldsymbol{r}\|_\infty \|\boldsymbol{w}\|_1 \geq -\eta\|\boldsymbol{w}\|_1.$$

- Claim: Lower bound of $\boldsymbol{w}^T\boldsymbol{r}$ is attained when $\boldsymbol{r} = -\eta \cdot \text{sign}(\boldsymbol{w})$

$$\begin{aligned}\boldsymbol{w}^T\boldsymbol{r} &= -\eta\boldsymbol{w}^T\text{sign}(\boldsymbol{w}) \\ &= -\eta\sum_{i=1}^{d} w_i\text{sign}(w_i) \\ &= -\eta\sum_{i=1}^{d} |w_i| \\ &= -\eta\|\boldsymbol{w}\|_1.\end{aligned}$$

- So the solution is $\boldsymbol{r} = -\eta \cdot \text{sign}(\boldsymbol{w})$.

To Summarize the Attack

Theorem (Maximum Loss ℓ_∞ Attack of Two-Class Linear Classifier)

The max-loss ℓ_∞ norm attack for a two-class linear classifier, i.e.,

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \boldsymbol{w}^T\boldsymbol{x} + w_0 \quad \text{subject to} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\|_\infty \leq \eta.$$

is given by

$$\boldsymbol{x} = \boldsymbol{x}_0 - \eta \cdot \mathit{sign}(\boldsymbol{w}).$$

- Compare to minimum-distance attack:

$$\boldsymbol{x} = \boldsymbol{x}_0 - \left(\frac{\boldsymbol{w}^T\boldsymbol{x}_0 + w_0}{\|\boldsymbol{w}\|_1}\right) \cdot \text{sign}(\boldsymbol{w}).$$

- η is now a free variable. You need to pick.

FGSM (Goodfellow et al., NeurIPS 2014)

- Define training loss as

$$J(\boldsymbol{x},\ \boldsymbol{w}) = g_t(\boldsymbol{x}) - \max_{i\neq t}\{g_i(\boldsymbol{x})\}$$
$$= -\left(\max_{i\neq t}\{g_i(\boldsymbol{x})\} - g_t(\boldsymbol{x})\right).$$

- Then max-loss attack is

$$\underset{\boldsymbol{x}}{\text{maximize}} \quad J(\boldsymbol{x},\boldsymbol{w}) \quad \text{subject to} \quad \|\boldsymbol{x}-\boldsymbol{x}_0\|_\infty \le \eta.$$

- Training: Minimize $J(\boldsymbol{x},\boldsymbol{w})$ by finding a good $\boldsymbol{w}$.
- Attack: Maximize $J(\boldsymbol{x},\boldsymbol{w})$ by finding a nasty $\boldsymbol{x}$.
- For neural networks, $J(\boldsymbol{x},\boldsymbol{w})$ can be very general.

FGSM (Goodfellow et al., NeurIPS 2014)

- How to attack $J(\boldsymbol{x}, \boldsymbol{w})$?
- Linearize:

$$J(\boldsymbol{x};\, \boldsymbol{w}) = J(\boldsymbol{x}_0 + \boldsymbol{r};\, \boldsymbol{w}) \approx J(\boldsymbol{x}_0;\, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{r}.$$

- Then solve

$$\underset{\boldsymbol{r}}{\text{maximize}} \quad J(\boldsymbol{x}_0;\, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{r} \quad \text{subject to} \quad \|\boldsymbol{r}\|_\infty \le \eta$$

- Equivalent to

$$\underset{\boldsymbol{r}}{\text{minimize}} \quad \underbrace{-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{r}}_{\boldsymbol{w}^T \boldsymbol{r}} - \underbrace{J(\boldsymbol{x}_0;\, \boldsymbol{w})}_{w_0} \quad \text{subject to} \quad \|\boldsymbol{r}\|_\infty \le \eta$$

- Solution is

$$\boldsymbol{r} = \eta \cdot \text{sign}(-\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w}))$$

FGSM (Goodfellow et al., NeurIPS 2014)

Definition (Fast Gradient Sign Method (FGSM) by Goodfellow et al 2014)

Given a loss function $J(\boldsymbol{x}; \boldsymbol{w})$, the FGSM creates an attack $\boldsymbol{x}$ by

$$\boldsymbol{x} = \boldsymbol{x}_0 + \eta \cdot \mathrm{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})). \tag{2}$$

Corollary (FGSM as a Max-Loss Attack Problem)

The FGSM attack can be formulated as the optimization with $J(\boldsymbol{x}; \boldsymbol{w})$ being the loss function:

$$\underset{\boldsymbol{r}}{\mathit{maximize}} \;\; \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})^T \boldsymbol{r} + J(\boldsymbol{x}_0; \boldsymbol{w}) \;\; \mathit{subject\ to} \;\; \|\boldsymbol{r}\|_\infty \leq \eta,$$

of which the solution is given by

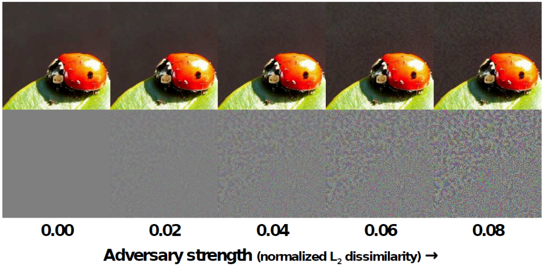
$$\boldsymbol{x} = \boldsymbol{x}_0 + \eta \cdot \mathit{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})). \tag{3}$$

FGSM (Goodfellow et al., NeurIPS 2014)

Definition (Fast Gradient Sign Method (FGSM) by Goodfellow et al 2014)

Given a loss function $J(\boldsymbol{x}; \boldsymbol{w})$, the FGSM creates an attack $\boldsymbol{x}$ by

$$\boldsymbol{x} = \boldsymbol{x}_0 + \eta \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})). \tag{4}$$



https://arxiv.org/pdf/1711.00117.pdf

ℓ_∞ and ℓ_2 FGSM

Corollary (FGSM as a Max-Loss Attack)

The FGSM attack can be formulated as the optimization with $J(\boldsymbol{x}; \boldsymbol{w})$ being the loss function:

$$\underset{\boldsymbol{r}}{\text{maximize}} \quad \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})^T \boldsymbol{r} + J(\boldsymbol{x}_0; \boldsymbol{w}) \quad \text{subject to} \quad \|\boldsymbol{r}\| \leq \eta,$$

of which the solution is given by

$$\boldsymbol{x} = \boldsymbol{x}_0 + \eta \cdot \text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})) \qquad (\ell_\infty\text{-norm})$$

and

$$\boldsymbol{x} = \boldsymbol{x}_0 + \eta \cdot \frac{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})}{\|\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0; \boldsymbol{w})\|_2} \qquad (\ell_2\text{-norm})$$

Iterative Fast Gradient Sign Method

- By Kurakin, Goodfellow and Bengio (ICLR 2017)
- Recall this equation

$$\begin{aligned} J(\boldsymbol{x};\, \boldsymbol{w}) &= J(\boldsymbol{x}_0 + \boldsymbol{r};\, \boldsymbol{w}) \\ &\approx J(\boldsymbol{x}_0;\, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{r} \\ &= J(\boldsymbol{x}_0;\, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T (\boldsymbol{x} - \boldsymbol{x}_0) \\ &= J(\boldsymbol{x}_0;\, \boldsymbol{w}) + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{x} - \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{x}_0. \end{aligned}$$

- Let us consider the problem

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{maximize}} \quad & \cancel{J(\boldsymbol{x}_0;\, \boldsymbol{w})} + \nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{x} - \cancel{\nabla_{\boldsymbol{x}} J(\boldsymbol{x}_0;\, \boldsymbol{w})^T \boldsymbol{x}_0} \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta, \;\; 0 \leq \boldsymbol{x} \leq 1. \end{aligned}$$

Iterative Gradient Sign Method

- Introduce iterative linearization

$$\mathbf{x}^{(k+1)} = \underset{\mathbf{x}}{\text{argmax}} \;\; \nabla_{\mathbf{x}} J(\mathbf{x}^{(k)};\, \mathbf{w})^T \mathbf{x} \quad \text{subject to} \;\; \|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \le \eta, \;\; 0 \le \mathbf{x} \le 1$$

- The optimization becomes

$$\begin{aligned}\mathbf{x}^{(k+1)} &= \underset{\mathbf{x}}{\text{argmax}} \;\; \nabla_{\mathbf{x}} J(\mathbf{x}^{(k)};\, \mathbf{w})^T \mathbf{x} \quad \text{subject to} \;\; \|\mathbf{x} - \mathbf{x}^{(k)}\|_\infty \le \eta, \;\; 0 \le \mathbf{x} \le 1 \\ &= \mathcal{P}_{[0,1]} \left\{ \mathbf{x}^{(k)} + \eta \cdot \text{sign}(\nabla_{\mathbf{x}} J(\mathbf{x}^{(k)};\, \mathbf{w})) \right\},\end{aligned}$$

- This is known as the projected gradient descent (PGD).
- Strongest first order attack, so far.
- You can add random noise to $\mathbf{x}^{(k)}$ to make it less predictable.

Outline

- Lecture 33 Overview
- Lecture 34 Min-distance attack
- Lecture 35 Max-loss attack and regularized attack

Today's Lecture

- Max-loss attack
  - Linear models
  - Deep models: FGSM and PGD
- Regularized attack
  - Linear models
  - CW attack

Two-Class Linear Classifier

- We want to study

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + \lambda\left(\max_{j\neq t}\{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x})\right).$$

- If we restrict to two-class, linear classifier, then simplified to

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + \lambda\left((\boldsymbol{w}_j^T\boldsymbol{x} + w_{j,0}) - (\boldsymbol{w}_t^T\boldsymbol{x} + w_{t,0})\right),$$

  which is

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + \lambda(\boldsymbol{w}^T\boldsymbol{x} + w_0).$$

- Unconstrained minimization.
- Let $\varphi(\boldsymbol{x}) = \frac{1}{2}\|\boldsymbol{x}-\boldsymbol{x}_0\|^2 + \lambda(\boldsymbol{w}^T\boldsymbol{x} + w_0)$. Then

$$0 = \nabla\varphi(\boldsymbol{x}) = (\boldsymbol{x}-\boldsymbol{x}_0) + \lambda\boldsymbol{w}.$$

- Solution is $\boldsymbol{x} = \boldsymbol{x}_0 - \lambda\boldsymbol{w}$.

Two-Class Linear Classifier

Theorem (Regularization-based Attack for Two-Class Linear Classifier)

The regularization-based attack for a two-class linear classifier generates the attack by solving

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda(\boldsymbol{w}^T\boldsymbol{x} + w_0),$$

of which the solution is given by

$$\boldsymbol{x} = \boldsymbol{x}_0 - \lambda\boldsymbol{w}.$$

- $\boldsymbol{w}$ is search direction
- λ is step size
- You need to choose λ.

Unboundedness of ℓ_1 Attack

- Can we do ℓ_1 attack?

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\|_1 + \lambda(\boldsymbol{w}^T\boldsymbol{x} + w_0),$$

which is equivalent to

$$\underset{\boldsymbol{r}}{\text{minimize}} \quad \|\boldsymbol{r}\|_1 + \lambda \boldsymbol{w}^T\boldsymbol{r}.$$

- The optimality condition is (sort of):

$$\text{sign}(r_i) + \lambda w_i = 0.$$

- This requires that

$$\lambda w_i = \begin{cases} \pm 1, & |r_i| > 0, \\ \in (-1, 1) & r_i = 0. \end{cases}$$

- So $|\lambda w_i|$ will never exceed 1.

Unboundedness of ℓ_1 Attack

- $$\lambda w_i = \begin{cases} \pm 1, & |r_i| > 0, \\ \in (-1, 1) & r_i = 0. \end{cases}$$
- Therefore, if $|\lambda \boldsymbol{w}| > \mathbf{1}$, then the above equation is impossible to hold regardless of how we choose $\boldsymbol{r}$.
- This means that the optimization does not have a solution.
- You can show that the function

  $$f(x) = |x| + \alpha x$$

  goes to $-\infty$ as $x \to -\infty$ if $\alpha > 1$.
- and goes to $-\infty$ as $x \to +\infty$ if $\alpha > -1$.
- So unbounded below.

Carlini-Wagner Attack (2016)

- The idea is to solve

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \lambda \cdot \max\left\{ \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right), 0 \right\},$$

- If $(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x})) < 0$: Already misclassified. No action needed.
- If $(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x})) > 0$: Not yet misclassified. Need action.
- Here we used the rectifier function

$$\zeta(x) = \max(x, 0).$$

- So the problem can be written as

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \lambda \cdot \zeta \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right).$$

- ¿7-¿ The norm here can be ℓ_1 or ℓ_2, or any other norm.

Comparing Regularized and Min-Norm

- Regularized attack is

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \lambda \cdot \zeta \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right).$$

- Min-distance attack is

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \iota_\Omega(\boldsymbol{x}),$$

where

$$\iota_\Omega(\boldsymbol{x}) = \begin{cases} 0, & \text{if} \quad \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \leq 0, \\ +\infty, & \text{otherwise.} \end{cases}$$

- So the regularized attack (CW attack) is a soft-version of the min-distance attack.

CW Attack for ℓ_1-norm

- We showed that this problem is unbounded below.

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\|_1 + \lambda(\boldsymbol{w}^T \boldsymbol{x} + w_0),$$

- Now consider the CW attack:

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\|_1 + \lambda \max\left(\boldsymbol{w}^T \boldsymbol{x} + w_0, 0\right).$$

- The objective function is always non-negative: $\|\boldsymbol{x} - \boldsymbol{x}_0\|_1 \geq 0$ and $\max\left(\boldsymbol{w}^T \boldsymbol{x} + w_0, 0\right) \geq 0$.
- We are guaranteed to have a solution.
- Here is a trivial solution.
- Lower bound is achieved when $\boldsymbol{x} = \boldsymbol{x}_0$ and $\boldsymbol{w}^T \boldsymbol{x}_0 + w_0 = 0$.
- This happens when the attack solution is $\boldsymbol{x} = \boldsymbol{x}_0$ and $\boldsymbol{x}_0$ is on the decision boundary.
- Of course, the chance for this to happen is unlikely. So we can safely ignore this trivial case.

Convexity for Linear Classifier

- The function $h(\boldsymbol{x}) = \max(\varphi(\boldsymbol{x}), 0)$ is convex in $\boldsymbol{x}$ if $\varphi(\boldsymbol{x})$ is convex.
-

$$
\begin{aligned}
h(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}) &= \max\left(\varphi(\alpha \boldsymbol{x} + (1-\alpha)\boldsymbol{y}), 0\right) \\
&\le \max\left(\alpha\varphi(\boldsymbol{x}) + (1-\alpha)\varphi(\boldsymbol{y}), 0\right) \\
&\le \alpha \max\left(\varphi(\boldsymbol{x}), 0\right) + (1-\alpha)\max(\varphi(\boldsymbol{y}), 0) \\
&= \alpha h(\boldsymbol{x}) + (1-\alpha) h(\boldsymbol{y}).
\end{aligned}
$$

- Our $\varphi(\boldsymbol{x}) = \boldsymbol{w}^T\boldsymbol{x} + w_0$. So φ is convex.
- So the overall optimization is convex

$$
\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \lambda \max\left(\boldsymbol{w}^T\boldsymbol{x} + w_0, 0\right).
$$

- That means you can solve using CVX.

General g

- In general, CW attack solves

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot \zeta \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right).$$

- We can use gradient algorithms.
- The gradient of $\zeta(\cdot)$ is

$$\frac{d}{ds}\zeta(s) = \mathbb{I}\{s > 0\} \stackrel{\text{def}}{=} \begin{cases} 1, & \text{if} \quad s > 0, \\ 0, & \text{otherwise.} \end{cases}$$

- Let $i^*(\boldsymbol{x})$ be the index of the maximum response
- $i^*(\boldsymbol{x}) = \underset{j \neq t}{\text{argmax}} \{g_j(\boldsymbol{x})\}$
- For the time being, let us assume that the index i^* is independent of $\boldsymbol{x}$
- Then, the gradient is

CW Attack Algorithm

- The gradient is

$$
\begin{aligned}
\nabla_{\boldsymbol{x}} \zeta &\left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right) \\
&= \nabla_{\boldsymbol{x}} \zeta \left(\{g_{i^*}(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right) \\
&= \begin{cases} \nabla_{\boldsymbol{x}} g_{i^*}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} g_t(\boldsymbol{x}), & \text{if} \quad g_{i^*}(\boldsymbol{x}) - g_t(\boldsymbol{x}) > 0, \\ 0, & \text{otherwise.} \end{cases} \\
&= \mathbb{I} \left\{ g_{i^*}(\boldsymbol{x}) - g_t(\boldsymbol{x}) > 0 \right\} \cdot \left(\nabla_{\boldsymbol{x}} g_{i^*}(\boldsymbol{x}) - \nabla_{\boldsymbol{x}} g_j(\boldsymbol{x}) \right)
\end{aligned}
$$

- Letting $\varphi(\boldsymbol{x})$ be the overall objective function

$$
\varphi(\boldsymbol{x}) = \|\boldsymbol{x} - \boldsymbol{x}_0\|^2 + \lambda \cdot \max \left\{ \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \right), 0 \right\},
$$

- The gradient is

$$
\nabla \varphi(\boldsymbol{x};\ i^*) = 2(\boldsymbol{x} - \boldsymbol{x}_0) + \lambda \cdot \mathbb{I} \left\{ g_{i^*}(\boldsymbol{x}) - g_t(\boldsymbol{x}) > 0 \right\} \cdot \left(\nabla g_{i^*}(\boldsymbol{x}) - \nabla g_j(\boldsymbol{x}) \right).
$$

CW Attack Algorithm

- Gradient is

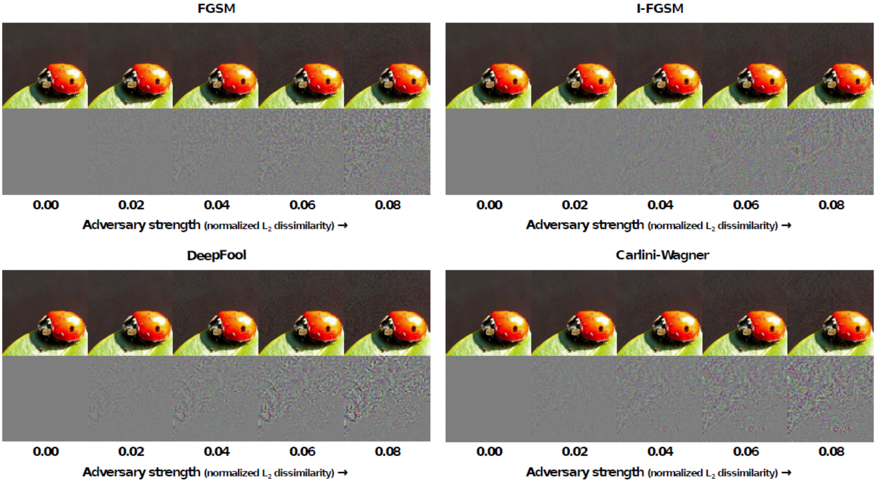
$$\nabla\varphi(\boldsymbol{x};\ i^*) = 2(\boldsymbol{x}-\boldsymbol{x}_0)+\lambda\cdot\mathbb{I}\left\{g_{i^*}(\boldsymbol{x}) - g_t(\boldsymbol{x}) > 0\right\}\cdot(\nabla g_{i^*}(\boldsymbol{x}) - \nabla g_j(\boldsymbol{x}))\,.$$

- The algorithm is
- For iteration $k = 1, 2, \ldots$

$$\begin{aligned} i^* &= \underset{j\neq t}{\text{argmax}}\ \{g_j(\boldsymbol{x}^k)\} \\ \boldsymbol{x}^{k+1} &= \boldsymbol{x}^k - \alpha\nabla\varphi(\boldsymbol{x}^k;\ i^*). \end{aligned}$$

- α is gradient descent step size. You need to tune it.
- λ is regularization parameter. You need to tune it.

Comparison



https://arxiv.org/pdf/1711.00117.pdf

Summary

So we have discussed three forms of adversarial attacks.

Min-Distance Attack

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{minimize}} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \\ \text{subject to} \quad & \max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x}) \leq 0, \end{aligned}$$

Max-Loss Attack

$$\begin{aligned} \underset{\boldsymbol{x}}{\text{maximize}} \quad & g_t(\boldsymbol{x}) - \max_{j \neq t} \{g_j(\boldsymbol{x})\} \\ \text{subject to} \quad & \|\boldsymbol{x} - \boldsymbol{x}_0\| \leq \eta, \end{aligned}$$

Regularized Attack

$$\underset{\boldsymbol{x}}{\text{minimize}} \quad \|\boldsymbol{x} - \boldsymbol{x}_0\| + \lambda \left(\max_{j \neq t} \{g_j(\boldsymbol{x})\} - g_t(\boldsymbol{x})\right)$$

- Next time we will talk about defense
- And then we will talk about fundamental trade off between robustness and accuracy