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 $x$ by solving the optimization

$$\begin{align*}
\text{maximize} & \quad g_t(x) - \max_{j \neq t} \{ g_j(x) \} \\
\text{subject to} & \quad \| x - x_0 \| \leq \eta,
\end{align*}$$

(1)

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

- I want to bound my attack $\| x - x_0 \| \leq \eta$
- I want to make $g_t(x)$ as big as possible
- So I want to maximize $g_t(x) - \max_{j \neq t} \{ g_j(x) \}$
- This is equivalent to

$$\begin{align*}
\text{minimize} & \quad \max_{j \neq t} \{ g_j(x) \} - g_t(x) \\
\text{subject to} & \quad \| x - x_0 \| \leq \eta,
\end{align*}$$
If you restrict yourself to two classes only ...

- The problem is

\[
\begin{align*}
\text{minimize} & \quad \max_{j \neq t} \left\{ g_j(x) \right\} - g_t(x) \\
\text{subject to} & \quad \|x - x_0\| \leq \eta,
\end{align*}
\]

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

\[
\begin{align*}
\text{minimize} & \quad w^T x + w_0 \quad \text{subject to} \quad \|x - x_0\| \leq \eta.
\end{align*}
\]

- Solvable in closed-form.
Max-Loss Attack using $\ell_2$-norm

- The problem is
  \[
  \min_{r} \ w^T r + b_0 \quad \text{subject to} \quad \|r\|_2 \leq \eta.
  \]
- Cauchy inequality:
  \[
  w^T r \geq -\|w\|_2 \|r\|_2 \geq -\eta \|w\|_2.
  \]

- Claim: Lower bound of $w^T r$ is attained when $r = -\eta w / \|w\|_2$:
  \[
  w^T r = w^T \left( -\frac{\eta w}{\|w\|_2} \right) = -\eta \|w\|_2.
  \]

- So the solution is $r = -\eta w / \|w\|_2$. 
Max-Loss Attack using $\ell_\infty$-norm

- Goal: Want to solve
  
  $$\min_{x} w^T x + w_0 \quad \text{subject to} \quad \|x - x_0\| \leq \eta.$$ 

- Define $x = x_0 + r$. Then
  
  $$w^T x + w_0 = w^T (x_0 + r) + w_0 = w^T x_0 + w^T r + w_0 = w^T r + w^T x_0 + w_0 = b_0$$

- Define $b_0 = (w^T x_0 + w_0)$. The optimization can be rewritten as
  
  $$\min_{r} w^T r + b_0 \quad \text{subject to} \quad \|r\|_\infty \leq \eta.$$
Solution to Max-Loss Attack (\(\ell_\infty\)-norm)

- Holder’s inequality (the negative side):
  \[
  \mathbf{w}^T \mathbf{r} \geq -\|\mathbf{r}\|_\infty \|\mathbf{w}\|_1 \geq -\eta \|\mathbf{w}\|_1.
  \]

- Claim: Lower bound of \(\mathbf{w}^T \mathbf{r}\) is attained when \(\mathbf{r} = -\eta \cdot \text{sign}(\mathbf{w})\)
  \[
  \mathbf{w}^T \mathbf{r} = -\eta \mathbf{w}^T \text{sign}(\mathbf{w})
  \]
  \[
  = -\eta \sum_{i=1}^{d} w_i \text{sign}(w_i)
  \]
  \[
  = -\eta \sum_{i=1}^{d} |w_i|
  \]
  \[
  = -\eta \|\mathbf{w}\|_1.
  \]

- So the solution is \(\mathbf{r} = -\eta \cdot \text{sign}(\mathbf{w})\).
To Summarize the Attack

Theorem (Maximum Loss $\ell_\infty$ Attack of Two-Class Linear Classifier)

The max-loss $\ell_\infty$ norm attack for a two-class linear classifier, i.e.,

$$\minimize_x \ w^T x + w_0 \quad \text{subject to} \quad \|x - x_0\|_\infty \leq \eta.$$ 

is given by

$$x = x_0 - \eta \cdot \text{sign}(w).$$

- Compare to minimum-distance attack:

$$x = x_0 - \left( \frac{w^T x_0 + w_0}{\|w\|_1} \right) \cdot \text{sign}(w).$$

- $\eta$ is now a free variable. You need to pick.
FGSM (Goodfellow et al., NeurIPS 2014)

- Define training loss as
  \[ J(x, w) = g_t(x) - \max_{i \neq t} \{ g_i(x) \} \]
  \[ = - \left( \max_{i \neq t} \{ g_i(x) \} - g_t(x) \right). \]

- Then max-loss attack is
  \[ \max_{x} J(x, w) \text{ subject to } \|x - x_0\|_\infty \leq \eta. \]

- Training: Minimize \( J(x, w) \) by finding a good \( w \).
- Attack: Maximize \( J(x, w) \) by finding a nasty \( x \).

- For neural networks, \( J(x, w) \) can be very general.
FGSM (Goodfellow et al., NeurIPS 2014)

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

  $$J(\mathbf{x}; \mathbf{w}) = J(\mathbf{x}_0 + \mathbf{r}; \mathbf{w}) \approx J(\mathbf{x}_0; \mathbf{w}) + \nabla_x J(\mathbf{x}_0; \mathbf{w})^T \mathbf{r}.$$ 

- Then solve

  $$\max_{\mathbf{r}} \ J(\mathbf{x}_0; \mathbf{w}) + \nabla_x J(\mathbf{x}_0; \mathbf{w})^T \mathbf{r} \quad \text{subject to} \quad \|\mathbf{r}\|_\infty \leq \eta$$

- Equivalent to

  $$\min_{\mathbf{r}} \ -\nabla_x J(\mathbf{x}_0; \mathbf{w})^T \mathbf{r} \ - \ J(\mathbf{x}_0; \mathbf{w}) \quad \text{subject to} \quad \|\mathbf{r}\|_\infty \leq \eta \quad \text{w}^T \mathbf{r} \ - \ w_0$$

- Solution is

  $$\mathbf{r} = \eta \cdot \text{sign}(-\nabla_x J(\mathbf{x}_0; \mathbf{w}))$$
FGSM (Goodfellow et al., NeurIPS 2014)

Definition (Fast Gradient Sign Method (FGSM) by Goodfellow et al 2014)
Given a loss function $J(x; w)$, the FGSM creates an attack $x$ by

$$x = x_0 + \eta \cdot \text{sign}(\nabla_x J(x_0; w)).$$  \hspace{1cm} (2)

Corollary (FGSM as a Max-Loss Attack Problem)
The FGSM attack can be formulated as the optimization with $J(x; w)$ being the loss function:

$$\max_r \quad \nabla_x J(x_0; w)^T r + J(x_0; w) \quad \text{subject to} \quad \|r\|_\infty \leq \eta,$$

of which the solution is given by

$$x = x_0 + \eta \cdot \text{sign}(\nabla_x J(x_0; w)).$$ \hspace{1cm} (3)
FGSM (Goodfellow et al., NeurIPS 2014)

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

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

$$x = x_0 + \eta \cdot \text{sign}(\nabla_x J(x_0; w)).$$

(4)

Corollary (FGSM as a Max-Loss Attack)

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

\[
\max_r \quad \nabla_x J(x_0; w)^T r + J(x_0; w) \quad \text{subject to} \quad \|r\| \leq \eta,
\]

of which the solution is given by

\[
x = x_0 + \eta \cdot \text{sign}(\nabla_x J(x_0; w)) \quad (\ell_\infty\text{-norm})
\]

and

\[
x = x_0 + \eta \cdot \frac{\nabla_x J(x_0; w)}{\|\nabla_x J(x_0; w)\|_2} \quad (\ell_2\text{-norm})
\]
Iterative Fast Gradient Sign Method

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

\[ J(x; w) = J(x_0 + r; w) \]
\[ \approx J(x_0; w) + \nabla_x J(x_0; w)^T r \]
\[ = J(x_0; w) + \nabla_x J(x_0; w)^T (x - x_0) \]
\[ = J(x_0; w) + \nabla_x J(x_0; w)^T x - \nabla_x J(x_0; w)^T x_0. \]

- Let us consider the problem

\[
\max_{x} \quad J(x_0; w) + \nabla_x J(x_0; w)^T x - \nabla_x J(x_0; w)^T x_0
\]

subject to \( \|x - x_0\| \leq \eta, \ 0 \leq x \leq 1. \)
Iterative Gradient Sign Method

- Introduce iterative linearization

\[ x^{(k+1)} = \arg\max_x \nabla_x J(x^{(k)}; w)^T x \]

subject to \( \|x - x^{(k)}\|_\infty \leq \eta, \ 0 \leq x \leq 1 \)

- The optimization becomes

\[ x^{(k+1)} = \arg\max_x \nabla_x J(x^{(k)}; w)^T x \]

subject to \( \|x - x^{(k)}\|_\infty \leq \eta, \ 0 \leq x \leq 1 \)

\[ = \mathcal{P}_{[0,1]} \left\{ x^{(k)} + \eta \cdot \text{sign}(\nabla_x J(x^{(k)}; w)) \right\}, \]

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

\[ \min_{x} \| x - x_0 \|^2 + \lambda \left( \max_{j \neq t} \{ g_j(x) \} - g_t(x) \right). \]

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

\[ \min_{x} \| x - x_0 \|^2 + \lambda \left( (w_j^T x + w_{j,0}) - (w_t^T x + w_{t,0}) \right), \]

which is

\[ \min_{x} \| x - x_0 \|^2 + \lambda (w^T x + w_0). \]

- Unconstrained minimization.
- Let \( \phi(x) = \frac{1}{2} \| x - x_0 \|^2 + \lambda (w^T x + w_0). \) Then

\[ 0 = \nabla \phi(x) = (x - x_0) + \lambda w. \]

- Solution is \( x = x_0 - \lambda w. \)
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

\[
\min_{\mathbf{x}} \frac{1}{2} \| \mathbf{x} - \mathbf{x}_0 \|^2 + \lambda (\mathbf{w}^T \mathbf{x} + w_0),
\]

of which the solution is given by

\[
\mathbf{x} = \mathbf{x}_0 - \lambda \mathbf{w}.
\]

- \( \mathbf{w} \) is search direction
- \( \lambda \) is step size
- You need to choose \( \lambda \).
Unboundedness of $\ell_1$ Attack

- Can we do $\ell_1$ attack?

$$\min_{x} \|x - x_0\|_1 + \lambda (w^T x + w_0),$$

which is equivalent to

$$\min_{r} \|r\|_1 + \lambda w^T r.$$

- 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 $\ell_1$ Attack

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

Therefore, if $|\lambda w| > 1$, then the above equation is impossible to hold regardless of how we choose $r$.

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

\[
\begin{align*}
\min_x \|x - x_0\| + \lambda \cdot \max \left\{ \left( \max_{j \neq t} \{ g_j(x) \} - g_t(x) \right), 0 \right\},
\end{align*}
\]

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

So the problem can be written as

\[
\begin{align*}
\min_x \|x - x_0\| + \lambda \cdot \zeta \left( \max_{j \neq t} \{ g_j(x) \} - g_t(x) \right).
\end{align*}
\]

The norm here can be \( \ell_1 \) or \( \ell_2 \), or any other norm.
Comparing Regularized and Min-Norm

- Regularized attack is
  \[
  \begin{aligned}
  \text{minimize } & \quad \|x - x_0\| + \lambda \cdot \zeta \left( \max_{j \neq t} \{g_j(x)\} - g_t(x) \right) \\
  \end{aligned}
  \]

- Min-distance attack is
  \[
  \begin{aligned}
  \text{minimize } & \quad \|x - x_0\| + \iota_\Omega(x),
  \end{aligned}
  \]

where

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

- So the regularized attack (CW attack) is a soft-version of the min-distance attack.
CW Attack for $\ell_1$-norm

- We showed that this problem is unbounded below.

$$\min_{x} \|x - x_0\|_1 + \lambda(w^T x + w_0),$$

- Now consider the CW attack:

$$\min_{x} \|x - x_0\|_1 + \lambda \max \left(w^T x + w_0, 0\right).$$

- The objective function is always non-negative: $\|x - x_0\|_1 \geq 0$ and $\max \left(w^T x + w_0, 0\right) \geq 0$.

- We are guaranteed to have a solution.

- Here is a trivial solution.

- Lower bound is achieved when $x = x_0$ and $w^T x_0 + w_0 = 0$.

- This happens when the attack solution is $x = x_0$ and $x_0$ 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 \( h(x) = \max(\varphi(x), 0) \) is convex in \( x \) if \( \varphi(x) \) is convex.

- \[
  h(\alpha x + (1 - \alpha)y) = \max(\varphi(\alpha x + (1 - \alpha)y), 0) \\
  \leq \max(\alpha \varphi(x) + (1 - \alpha)\varphi(y), 0) \\
  \leq \alpha \max(\varphi(x), 0) + (1 - \alpha) \max(\varphi(y), 0) \\
  = \alpha h(x) + (1 - \alpha) h(y).
\]

- Our \( \varphi(x) = w^T x + w_0 \). So \( \varphi \) is convex.

- So the overall optimization is convex

  \[
  \min_x \|x - x_0\| + \lambda \max \left( w^T x + w_0, 0 \right).
  \]

- That means you can solve using CVX.
General $g$

- In general, CW attack solves

$$\min_{\mathbf{x}} \| \mathbf{x} - \mathbf{x}_0 \|^2 + \lambda \cdot \zeta \left( \max_{j \neq t} \{ g_j(\mathbf{x}) \} - g_t(\mathbf{x}) \right).$$

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

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

- Let $i^*(\mathbf{x})$ be the index of the maximum response

$$i^*(\mathbf{x}) = \arg\max_{j \neq t} \{ g_j(\mathbf{x}) \}$$

- For the time being, let us assume that the index $i^*$ is independent of $\mathbf{x}$
- Then, the gradient is
CW Attack Algorithm

- The gradient is
  \[
  \nabla_x \zeta \left( \max_{j \neq t} \{ g_j(x) \} - g_t(x) \right)
  = \nabla_x \zeta \left( \{ g_i^*(x) \} - g_t(x) \right)
  = \begin{cases} 
    \nabla_x g_i^*(x) - \nabla_x g_t(x), & \text{if } g_i^*(x) - g_t(x) > 0, \\
    0, & \text{otherwise.}
  \end{cases}
  = \mathbb{I} \{ g_i^*(x) - g_t(x) > 0 \} \cdot (\nabla_x g_i^*(x) - \nabla_x g_j(x))
  \]

- Letting \( \varphi(x) \) be the overall objective function
  \[
  \varphi(x) = \|x - x_0\|^2 + \lambda \cdot \max \left\{ \left( \max_{j \neq t} \{ g_j(x) \} - g_t(x) \right), 0 \right\},
  \]

- The gradient is
  \[
  \nabla \varphi(x; i^*) = 2(x - x_0) + \lambda \cdot \mathbb{I} \{ g_i^*(x) - g_t(x) > 0 \} \cdot (\nabla g_i^*(x) - \nabla g_j(x)).
  \]
CW Attack Algorithm

- Gradient is

\[ \nabla \varphi(x; i^*) = 2(x-x_0) + \lambda \cdot \mathbb{I} \{ g_{i^*}(x) - g_t(x) > 0 \} \cdot (\nabla g_{i^*}(x) - \nabla g_j(x)) \]

- The algorithm is

For iteration \( k = 1, 2, \ldots \)

\[ i^* = \arg \max_{j \neq t} \{ g_j(x^k) \} \]

\[ x^{k+1} = x^k - \alpha \nabla \varphi(x^k; i^*) \]

- \( \alpha \) is gradient descent step size. You need to tune it.

- \( \lambda \) is regularization parameter. You need to tune it.
Comparison

Summary

So we have discussed three forms of adversarial attacks.

Min-Distance Attack

\[
\begin{align*}
\text{minimize} & \quad \|x - x_0\| \\
\text{subject to} & \quad \max_{j \neq t} \{ g_j(x) \} - g_t(x) \leq 0,
\end{align*}
\]

Max-Loss Attack

\[
\begin{align*}
\text{maximize} & \quad g_t(x) - \max_{j \neq t} \{ g_j(x) \} \\
\text{subject to} & \quad \|x - x_0\| \leq \eta,
\end{align*}
\]

Regularized Attack

\[
\begin{align*}
\text{minimize} & \quad \|x - x_0\| + \lambda \left( \max_{j \neq t} \{ g_j(x) \} - g_t(x) \right)
\end{align*}
\]

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