

ECE 595: Homework 5

Yi Qiao, Class ID 187
(Spring 2019)

Exercise 1: Adversarial Attacks on Gaussian Classifier

(a) minimum-norm attacks

(i) minimum l_2 and l_∞ attack

Since we only have 2 classes, the question becomes

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{x} - \mathbf{x}_0\| \\ & \text{subject to} \quad \mathbf{w}^T \mathbf{x} + w_0 = 0 \end{aligned} \tag{1}$$

using l_2 norm

the problem is the same as

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \\ & \text{subject to} \quad \mathbf{w}^T \mathbf{x} + w_0 = 0 \end{aligned} \tag{2}$$

The lagrangian is

$$\mathcal{L}(\mathbf{x}, \lambda) = \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2 + \lambda(\mathbf{w}^T \mathbf{x} + w_0) \tag{3}$$

Taking the derivative

$$\begin{aligned} \nabla_{\mathbf{x}} \mathcal{L}(\mathbf{x}, \lambda) &= \mathbf{x} - \mathbf{x}_0 + \lambda \mathbf{w} = 0 \\ \frac{\partial}{\partial \lambda} \mathcal{L}(\mathbf{x}, \lambda) &= \mathbf{w}^T \mathbf{x} + w_0 = 0 \end{aligned} \tag{4}$$

$$\begin{aligned} \lambda \mathbf{w} &= \mathbf{x}_0 - \mathbf{x} \\ \lambda \mathbf{w}^T \mathbf{w} &= \mathbf{w}^T \mathbf{x}_0 - \mathbf{w}^T \mathbf{x} \end{aligned} \tag{5}$$

$$\lambda = (\mathbf{w}^T \mathbf{w})^{-1} (\mathbf{w}^T \mathbf{x}_0 + w_0)$$

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_0 - \lambda \mathbf{w} \\ &= \mathbf{x}_0 - \frac{\mathbf{w}(\mathbf{w}^T \mathbf{x}_0 + w_0)}{\|\mathbf{w}\|_2^2} \end{aligned} \tag{6}$$

using l_∞ norm

$$\begin{aligned} & \underset{\mathbf{x}}{\text{minimize}} \quad \|\mathbf{x} - \mathbf{x}_0\|_\infty \\ & \text{subject to} \quad \mathbf{w}^T \mathbf{x} + w_0 = 0 \end{aligned} \tag{7}$$

Let $\mathbf{r} = \mathbf{x} - \mathbf{x}_0$, $b_0 = -(\mathbf{w}^T \mathbf{x}_0 + w_0)$, the problem becomes:

$$\begin{aligned} & \underset{\mathbf{x}}{\text{argmin}} \quad \|\mathbf{x} - \mathbf{x}_0\|_\infty \\ & \text{subject to} \quad \mathbf{w}^T \mathbf{r} = b_0 \end{aligned} \tag{8}$$

The lagrangian is

$$\mathcal{L}(\mathbf{r}, \lambda) = \|\mathbf{r}\|_\infty + \lambda(b_0 - \mathbf{w}^T \mathbf{r}) \quad (9)$$

Taking derivative,

$$\frac{\partial}{\partial \lambda} \mathcal{L}(\mathbf{r}, \lambda) = b_0 - \mathbf{w}^T \mathbf{r} = 0 \quad (10)$$

By Holder's Inequality:

$$\begin{aligned} |b_0| &= |\mathbf{w}^T \mathbf{r}| \leq \|\mathbf{w}\|_1 \|\mathbf{r}\|_\infty \\ \|\mathbf{r}\|_\infty &\geq \frac{|b_0|}{\|\mathbf{w}\|_1} \end{aligned} \quad (11)$$

Consider $\mathbf{r} = \eta \cdot \text{sign}(\mathbf{w})$, for some constant η tbd. We can show that

$$\|\mathbf{r}\|_\infty = \underset{i}{\operatorname{argmax}} |\eta \cdot \text{sign}(w_i)| = |\eta| \quad (12)$$

let $\eta = \frac{b_0}{\|\mathbf{w}\|_1} \cdot \text{sign}(\mathbf{w})$, then we have,

$$\|\mathbf{r}\|_\infty = |\eta| = \frac{b_0}{\|\mathbf{w}\|_1} \quad (13)$$

Lower bound is achieved, thus the solution is,

$$\mathbf{r} = \frac{|b_0|}{\|\mathbf{w}\|_1} \cdot \text{sign}(\mathbf{w}) \quad (14)$$

(ii) DeepFool attack

$$\begin{aligned} &\underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \\ &\text{subject to } g(\mathbf{x}) = 0 \end{aligned} \quad (15)$$

First order approximation

$$g(\mathbf{x}) \approx g(\mathbf{x}^{(k)}) + \nabla_{\mathbf{x}} g(\mathbf{x}^{(k)})^T (\mathbf{x} - \mathbf{x}^{(k)}) \quad (16)$$

Then the problem can be approximate by

$$\begin{aligned} &\underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \\ &\text{subject to } g(\mathbf{x}^{(k)}) + \nabla_{\mathbf{x}} g(\mathbf{x}^{(k)})^T (\mathbf{x} - \mathbf{x}^{(k)}) = 0 \end{aligned} \quad (17)$$

Let $\mathbf{w}^{(k)} = \nabla_{\mathbf{x}} g(\mathbf{x}^{(k)})$ and $w_0^{(k)} = g(\mathbf{x}^{(k)}) - \nabla_{\mathbf{x}} g(\mathbf{x}^{(k)})^T \mathbf{x}^{(k)}$

Then the problem is equivalent to

$$\begin{aligned} &\underset{\mathbf{x}}{\operatorname{argmin}} \|\mathbf{x} - \mathbf{x}_0\|_2^2 \\ &\text{subject to } (\mathbf{w}^{(k)})^T \mathbf{x} + w_0^{(k)} = 0 \end{aligned} \quad (18)$$

This is the same problem as minimum l_2 norm attack, Thus the solution will be,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \frac{((\mathbf{w}^{(k)})^T \mathbf{x}^{(k)} + w_0^{(k)}) \mathbf{w}^{(k)}}{\|\mathbf{w}^{(k)}\|_2^2} \quad (19)$$

substitute \mathbf{w} and w_0 back, we get

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \left(\frac{g(\mathbf{x}^{(k)})}{\|\nabla_{\mathbf{x}} g(\mathbf{x}^{(k)})\|^2} \right) \nabla_{\mathbf{x}} g(\mathbf{x}^{(k)}) \quad (20)$$

(iii) An example DeepFool never converge

something

(b) maximum-allowable attack

(i) l_∞ attack in the linear case

The problem is,

$$\begin{aligned} \underset{\mathbf{x}}{\operatorname{argmin}} \quad & \mathbf{w}^T \mathbf{x} + w_0 \\ \text{subject to} \quad & \|\mathbf{x} - \mathbf{x}_0\|_\infty < \eta \end{aligned} \quad (21)$$

let $\mathbf{x} = \mathbf{x}_0 + \mathbf{r}$, $b_0 = (\mathbf{w}^T \mathbf{x}_0 + w_0)$, the problem becomes,

$$\begin{aligned} \underset{\mathbf{r}}{\operatorname{argmin}} \quad & \mathbf{w}^T \mathbf{r} + b_0 \\ \text{subject to} \quad & \|\mathbf{r}\|_\infty < \eta \end{aligned} \quad (22)$$

by Holder's inequality,

$$\mathbf{w}^T \mathbf{r} \geq -\|\mathbf{r}\|_\infty \|\mathbf{w}\|_1 \geq -\eta \|\mathbf{w}\|_1 \quad (23)$$

as shown in the lecture note, the solution

$$\mathbf{r} = -\eta \cdot \operatorname{sign}(\mathbf{w}) \quad (24)$$

(ii) FGSM attack

$$\begin{aligned} \underset{\mathbf{x}}{\operatorname{argmax}} \quad & J(\mathbf{x}, \mathbf{w}) \\ \text{subject to} \quad & \|\mathbf{x} - \mathbf{x}_0\|_\infty \leq \eta \end{aligned} \quad (25)$$

Approximately, $J(\mathbf{x}, \mathbf{w}) = J(\mathbf{x}_0 + \mathbf{r}, \mathbf{w}) \approx J(\mathbf{x}_0, \mathbf{w}) + \nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})^T \mathbf{r}$

Then, the problem becomes

$$\begin{aligned} \underset{\mathbf{x}}{\operatorname{argmin}} \quad & -J(\mathbf{x}_0, \mathbf{w}) - \nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})^T \mathbf{r} \\ \text{subject to} \quad & \|\mathbf{r}\|_\infty \leq \eta \end{aligned} \quad (26)$$

of which the solution is given by

$$\mathbf{x} = \mathbf{x}_0 + \eta \cdot (\nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})) \quad (27)$$

In the problem setup, we get $J(\mathbf{x}) = -g(\mathbf{x})$, thus the solution is

$$\begin{aligned} \mathbf{x} &= \mathbf{x}_0 - \eta \cdot (\nabla_{\mathbf{x}} g(\mathbf{x}_0)) \\ &= \mathbf{x}_0 - \eta \cdot ((\mathbf{W}_j - \mathbf{W}_t) \mathbf{x}_0 + (\mathbf{w}_j - \mathbf{w}_t)) \\ &= (\eta(\mathbf{W}_t - \mathbf{W}_j) + 1) \mathbf{x}_0 + \eta(\mathbf{w}_t - \mathbf{w}_j) \end{aligned} \quad (28)$$

(iii) I-FGSM attack

$$\begin{aligned}
J(\mathbf{x}, \mathbf{w}) &= J(\mathbf{x}_0 + \mathbf{r}, \mathbf{w}) \\
&\approx J(\mathbf{x}_0, \mathbf{w}) + \nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})^T \mathbf{r} \\
&= J(\mathbf{x}_0, \mathbf{w}) + \nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})^T (\mathbf{x} - \mathbf{x}_0) \\
&= J(\mathbf{x}_0, \mathbf{w}) + \nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})^T \mathbf{x} - \nabla_{\mathbf{x}} J(\mathbf{x}_0, \mathbf{w})^T \mathbf{x}_0
\end{aligned} \tag{29}$$

$$\begin{aligned}
\mathbf{x}^{(k+1)} &= \underset{0 \leq \mathbf{x} \leq 1}{\operatorname{argmax}} J(\mathbf{x}^{(k)}, \mathbf{w}) \text{ subject to } \|\mathbf{x} - \mathbf{x}_0\| \leq \eta \\
&= \underset{0 \leq \mathbf{x} \leq 1}{\operatorname{argmax}} \nabla_{\mathbf{x}} J(\mathbf{x}^{(k)}, \mathbf{w})^T \mathbf{x} \text{ subject to } \|\mathbf{x} - \mathbf{x}_0\| \leq \eta \\
&= \mathcal{P} \left\{ \mathbf{x}^{(k)} + \eta \cdot \operatorname{sign} \left(\nabla_{\mathbf{x}} J(\mathbf{x}^{(k)}, \mathbf{w}) \right) \right\}
\end{aligned} \tag{30}$$

(c) Regularization based attack

(i) linear case

$$\underset{\mathbf{x}}{\operatorname{argmin}} \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2 + \lambda(\mathbf{w}^T \mathbf{x} + w_0) \tag{31}$$

Taking derivative

$$\begin{aligned}
\nabla_{\mathbf{x}} \frac{1}{2} \|\mathbf{x} - \mathbf{x}_0\|_2^2 + \lambda(\mathbf{w}^T \mathbf{x} + w_0) \\
= \mathbf{x} - \mathbf{x}_0 + \lambda \mathbf{w} = 0
\end{aligned} \tag{32}$$

Solve for \mathbf{x} ,

$$\mathbf{x} = \mathbf{x}_0 - \lambda \mathbf{w} \tag{33}$$

(ii)

$$\begin{aligned}
&\underset{\mathbf{x}}{\operatorname{argmin}} \varphi(\mathbf{x}), \text{ where} \\
\varphi(\mathbf{x}) &= \|\mathbf{x} - \mathbf{x}_0\|_2^2 + \lambda \zeta(g_j(\mathbf{x}) - g_t(\mathbf{x})), \\
\zeta(y) &= \max(y, 0), \text{ and } j \neq t
\end{aligned} \tag{34}$$

Taking the derivative,

$$\nabla \varphi(\mathbf{x}) = 2(\mathbf{x} - \mathbf{x}_0) + \lambda \mathbb{I}\{g_j(\mathbf{x}) - g_t(\mathbf{x}) > 0\} \cdot (\nabla_{\mathbf{x}} g_j(\mathbf{x}) - \nabla_{\mathbf{x}} g_t(\mathbf{x})) \tag{35}$$

Using my favorite gradient descent, we can tell,

$$\mathbf{x}^{(k+1)} = \mathbf{x}^{(k)} - \alpha \nabla_{\mathbf{x}} \varphi(\mathbf{x}^{(k)}) \tag{36}$$

substituting in $g(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T (\mathbf{W}_j - \mathbf{W}_t) \mathbf{x} + (\mathbf{w}_j - \mathbf{w}_t)^T \mathbf{x} + (w_{j,0} - w_{t,0})$, we get

$$\begin{aligned}
\nabla_{\mathbf{x}} g(\mathbf{x}) &= (\mathbf{W}_j - \mathbf{W}_t) \mathbf{x} + (\mathbf{w}_j - \mathbf{w}_t) \\
\mathbf{x}^{(k+1)} &= \mathbf{x}^{(k)} - 2\alpha(\mathbf{x}^{(k)} - \mathbf{x}_0) - \alpha \lambda \mathbb{I}\{g(\mathbf{x}) > 0\} \cdot (\nabla_{\mathbf{x}} g(\mathbf{x})) \\
&= \mathbf{x}^{(k)} - 2\alpha(\mathbf{x}^{(k)} - \mathbf{x}_0) - \alpha \lambda \mathbb{I}\{g(\mathbf{x}) > 0\} \cdot ((\mathbf{W}_j - \mathbf{W}_t) \mathbf{x} + (\mathbf{w}_j - \mathbf{w}_t))
\end{aligned} \tag{37}$$

Exercise 2: CW-Attack on Gaussian Classifier - Non-overlapping Patches

(a)&(b)

refer to code in the back

(c)

(i) & (ii) The final perturbed images & their perturbation respectively



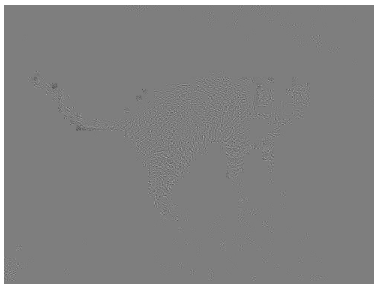
(a) $\lambda = 1$



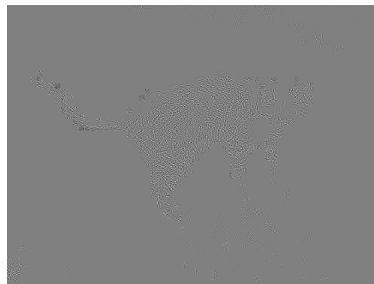
(b) $\lambda = 5$



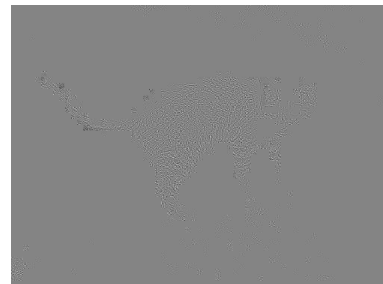
(c) $\lambda = 10$



(d) $\lambda = 1$



(e) $\lambda = 5$



(f) $\lambda = 10$

Figure 1: perturbed images and their perturbations with $\lambda = 1, 5, 10$

(iii) Frobenius norm of the perturbation

$\lambda = 1$, Frobenius norm: 8.253352243144604

$\lambda = 5$, Frobenius norm: 8.928742803780679

$\lambda = 10$, Frobenius norm: 9.697938770929634

(iv) The classifier's output

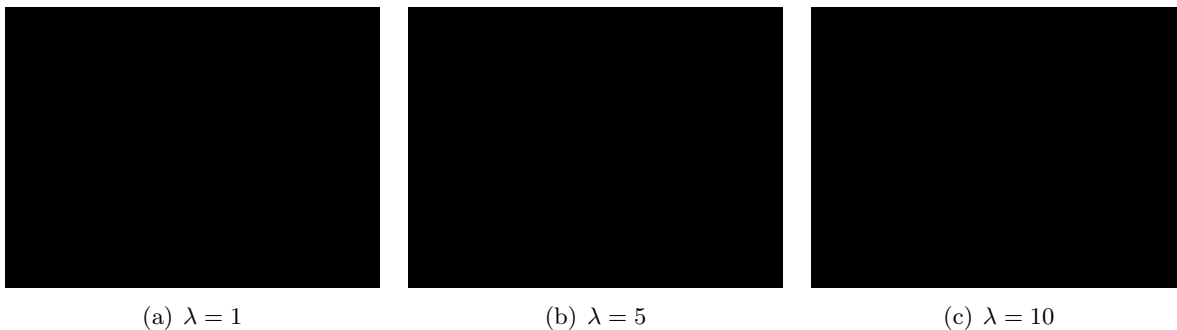


Figure 2: classified perturbed image with $\lambda = 1, 5, 10$

For all three cases, all patches are classified as grass after applying perturbation.

(v) Plots during gradient descent

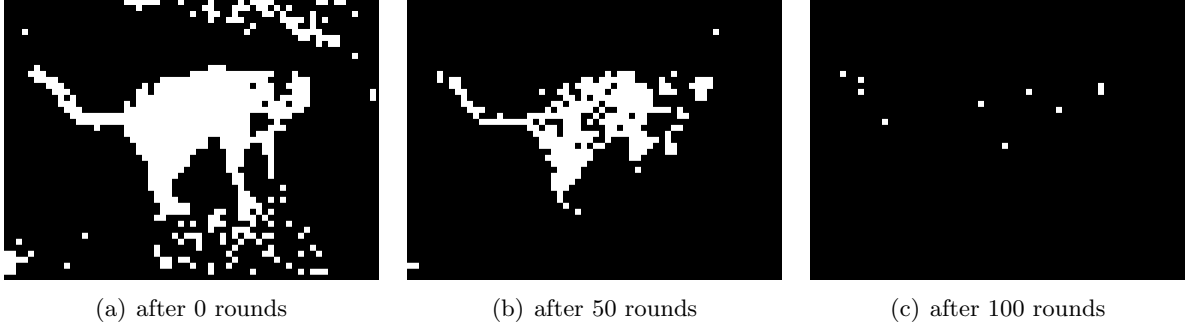


Figure 3: classified perturbed image during gradient descent with $\lambda = 1$

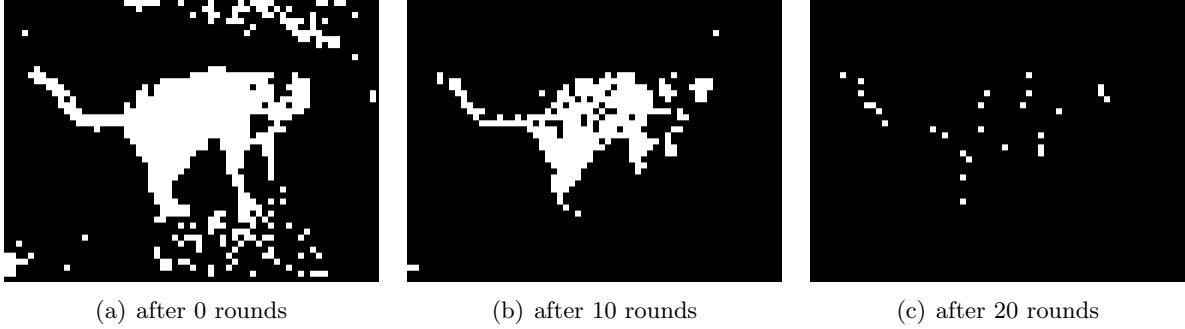


Figure 4: classified perturbed image during gradient descent with $\lambda = 5$

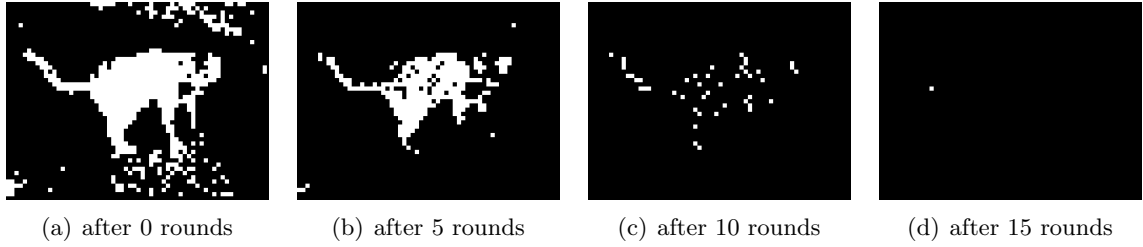


Figure 5: classified perturbed image during gradient descent with $\lambda = 10$

From the above plots, we can obviously see that increase λ will speed up the attack.

Exercise 3: CW Attack on Gaussian Classifier - Overlapping Patches

$$\begin{aligned}
 \mathbf{X}^* &= \underset{\mathbf{X}}{\operatorname{argmin}} \sum_{i=1}^L \{ \|\mathbf{P}_i(\mathbf{X} - \mathbf{X}_0)\|_2^2 + \lambda \max(g_j(\mathbf{P}_i\mathbf{X}) - g_t(\mathbf{P}_i\mathbf{X}), 0) \} \\
 &= \underset{\mathbf{X}}{\operatorname{argmin}} \|\mathbf{X} - \mathbf{X}_0\|_2^2 + \lambda \sum_{i=1}^L \max(g_j(\mathbf{P}_i\mathbf{X}) - g_t(\mathbf{P}_i\mathbf{X}), 0)
 \end{aligned} \tag{38}$$

(a)

the gradient of the above function is,

$$\nabla_{\mathbf{X}} = 2(\mathbf{X} - \mathbf{X}_0) + \lambda \mathbb{I}\{g_j(\mathbf{P}_i\mathbf{X}) - g_t(\mathbf{P}_i\mathbf{X}) > 0\} \times (\nabla_{\mathbf{X}}(g_j(\mathbf{P}_i\mathbf{X}) - g_t(\mathbf{P}_i\mathbf{X}))) \tag{39}$$