Homework Assignment 1

Response by: **TODO: Soham Roy (7028704)**

Due: **2:00pm Thursday, 23 November 2023 on CISPA CMS**

Collaboration Policy: You should do this assignment by yourself and submit your own answers. You may discuss the problems with anyone you want and it is also fine to get help from anyone on problems with LaTeX or Jupyter/Python. You should note in the *Collaborators* box below the people you collaborated with.

Collaborators: TODO: Subrat Kumar Dutta , Somrita Ghosh

Problem 1 (10 pts) Consider a binary logistic regression model with loss $\ell(\boldsymbol{w};\boldsymbol{x},y) = -\log\sigma(y\cdot\langle\boldsymbol{w},\boldsymbol{x}\rangle)$, where $\boldsymbol{x}\in\mathbb{R}^d$, $y\in\{-1,+1\}$, and $\sigma(z) = 1/(1+\exp(-z))$. Let $\mathcal{B}_\epsilon(\boldsymbol{x},\ell_\infty) = \{\boldsymbol{x}'\in\mathbb{R}^d : \|\boldsymbol{x}'-\boldsymbol{x}\|_\infty \leq \epsilon\}$ limit the searching region of feasible $\boldsymbol{x}'$. Show that

$$\max_{\boldsymbol{x}'\in\mathcal{B}_\epsilon(\boldsymbol{x},\ell_\infty)} \ell(\boldsymbol{w};\boldsymbol{x}',y) = -\log\sigma(y\cdot\langle\boldsymbol{w},\boldsymbol{x}\rangle - \epsilon\|\boldsymbol{w}\|_1).$$

we have $\|x-x'\| \leq \epsilon$ and $\|w,(x-x')\|_\infty \leq \epsilon\cdot\|w\|_1$

we can write $(w,x') = \langle w,x\rangle + w\cdot\langle x-x'\rangle$
l(w;x',y) = -log($\sigma\langle w,x'\rangle$)
$l(w;x',y) = -\log(\sigma(y\langle w,x\rangle + \langle w,x'-x\rangle))$
$l(w;x',y) = -\log(\sigma(\langle y,w\rangle + y\epsilon\|w\|_1))$
$\max l(w;x',y) = \min(y\langle w,x\rangle + y\epsilon\|w\|_1)$

now we know that $y\langle w,x\rangle + y\epsilon\|w\|_1 \geq y\langle w,x\rangle - y\epsilon\|w\|_1$ because the norm is always ≥ 0 and so to maximize l(w;x',y), the argument in the sigmoid function will be

$\max_{\boldsymbol{x}'\in\mathcal{B}_\epsilon(\boldsymbol{x},\ell_\infty)} \ell(\boldsymbol{w};\boldsymbol{x}',y) = -\log\sigma(y\cdot\langle\boldsymbol{w},\boldsymbol{x}\rangle - \epsilon\|\boldsymbol{w}\|_1).$

Problem 2 (10 pts) Suppose we want to solve the maximization problem of logistic regression specified in Problem 1 using gradient descent. Compute the gradient of $\ell(\boldsymbol{w};\boldsymbol{x},y)$ with respect to $\boldsymbol{x}$.

l(w;x,y) = -log($\sigma\langle w,x\rangle$)
$l(w;x,y) = -\log\left(\frac{1}{1+e^{-ywx}}\right)$
$\frac{\partial l(w;x,y)}{\partial x} = -yw\left(\frac{1}{1+e^{-ywx}}\right)$
$\frac{\partial l(w;x,y)}{\partial x} = -yw(1+\sigma(y\langle w,x\rangle))$

Implementation Problems. Below are two problems that you need to complete the provided Jupyter notebook. The goal is to help you understand the iterative PGD attack (both untargeted and targeted versions). For illustration, we will use a pretrained ImageNet ResNet50 model as the victim, and use a ladybug image from ImageNet as the seed example. Note that the class index of ladybug is 301.

Problem 3 (10 pts) Let K be the number of class labels. Consider ℓ_∞ perturbations with $\epsilon = 2/255$. The goal of PGD attack is to solve the following objective using an iterative algorithm:

$$\max_{\boldsymbol{x}'} \ \ell(h_\theta(\boldsymbol{x}', y)) \ \text{ subject to } \ \|\boldsymbol{x}' - \boldsymbol{x}\|_\infty \leq \epsilon,$$

where $(\boldsymbol{x}, y)$ is the input example (in this task, a ladybug image), $h_\theta : \mathbb{R}^d \to \mathbb{R}^K$ is neural network mapping from input to logit layer (in this task, the pretrained ResNet50 model), and ℓ is the cross-entropy loss.

Your Task: Write down the algorithm pseudocode of untargeted PGD attack, then implement the iterative attack by completing the corresponding section of the provided Jupyter notebook. Specifically, you need to initialize the PGD attack in the implementation with zero initialization, and run PGD attack using a SGD optimizer with a learning rate 0.1 for 30 iterations. In this case, you are using the raw gradient without the sign function, as described in the Note of Problem 2. Remember that you also need to ensure the output of your algorithm lies within $\mathcal{B}_\epsilon(\boldsymbol{x}, \ell_\infty)$.

Problem 4 (10 pts) Note that the previous implementation of PGD attack is untargeted, which does not specify a targeted label to guide the adversarial examples generation process. Under the same setting of Problem 3, the targeted version of PGD attack is designed to solve the following objective:

$$\max_{\boldsymbol{x}'} \ \Big(\ell(h_\theta(\boldsymbol{x}', y)) - \ell(h_\theta(\boldsymbol{x}', y_{\text{targ}}))\Big) \ \text{ subject to } \ \|\boldsymbol{x}' - \boldsymbol{x}\|_\infty \leq \epsilon,$$

where y_{targ} is a pre-selected targeted label that is different from y.

Your Task: Write down the algorithm pseudocode of targeted PGD attack, then implement the attack by completing the corresponding section of the provided Jupyter notebook. Specifically, the target label should be set as zebra (the corresponding class index is 340 in ImageNet). You need to initialize the PGD attack in the implementation with zero initialization, and run PGD attack using a SGD optimizer with a learning rate 0.005 for 100 iterations.

Problem 5 (bonus, 5 pts) Consider a linear model with soft-SVM loss $\ell(\boldsymbol{w}; \boldsymbol{x}, y) = \max(0, 1 - y \cdot \langle \boldsymbol{w}, \boldsymbol{x} \rangle)$, where $\boldsymbol{x} \in \mathbb{R}^d$ and $y \in \{-1, +1\}$. For any $p \geq 1$, show that

$$\max_{\boldsymbol{x}' \in \mathcal{B}_\epsilon(\boldsymbol{x}, \ell_p)} \ell(\boldsymbol{w}; \boldsymbol{x}', y) = \max(0, 1 - y \cdot \langle \boldsymbol{w}, \boldsymbol{x} \rangle + \epsilon \|\boldsymbol{w}\|_q),$$

where $\mathcal{B}_\epsilon(\boldsymbol{x}, \ell_p) = \{\boldsymbol{x}' \in \mathbb{R}^d : \|\boldsymbol{x}' - \boldsymbol{x}\|_p \leq \epsilon\}$ is the ϵ-ball at $\boldsymbol{x}$ in ℓ_p-norm, and q satisfies $1/p + 1/q = 1$.

we have $\|x - x'\|_p \leq \epsilon$
$and \|\langle w, x - x' \rangle\|_p \leq \|w\|_q \epsilon$ from Hölder's Theorem.
we can write $(w, x') = \langle w, x \rangle + \langle w, x - x' \rangle \leq \langle w, x \rangle \ + \|w\|_q \epsilon$

Putting the above in the objective function, we want to maximize $1 - y\langle w, x \rangle$.

so we have $(w, x') = \langle w, x \rangle + \langle w, x - x' \rangle \leq \langle w, x \rangle \ + \|w\|_q \epsilon$

Multiplying with -1 on both sides and adding +1 we get
1 - y $\langle w, x \rangle + \epsilon \|w\|_q \geq 1 - y\langle w, x \rangle - y\langle w, x' - x \rangle \geq 1 - y\langle w, x \rangle - y\epsilon \|w\|_q$

hence to maximize l we would have to $\max(0, 1 - y \cdot \langle \boldsymbol{w}, \boldsymbol{x} \rangle + \epsilon \|\boldsymbol{w}\|_q)$

End of Homework Assignment 1 (PDF part)
Don't forget to also complete and submit the Jupyter notebook!