The LogBarrier adversarial attack: making effective use of decision boundary information

Chris Finlay, Aram-Alexandre Pooladian & Adam M Oberman
Department of Mathematics & Statistics, McGill University, Montréal Canada
{christopher.finlay, aram-alexandre.pooladian}@mail.mcgill.ca, adam.oberman@mcgill.ca

McGill

Overview

- Adversarial attacks for image classification are small perturbations to images that are designed to cause misclassification by a model [1].
- Adversarial attacks formally correspond to an optimization problem: find a minimum norm image perturbation, constrained to cause misclassification

$$\begin{aligned} &\underset{\delta}{\text{minimize}} \quad \|\delta\| \\ &\text{subject to} \quad \arg\max f(x+\delta) \neq c, \end{aligned} \tag{1}$$

where $f(x)$ is the model's prediction, and c is the correct label.
- However, to date, no gradient-based attacks have used best practices from the optimization literature to solve this constrained minimization problem.
- We design a new untargeted attack, based on these best practices, using the well-regarded *logarithmic barrier method* [2].

The LogBarrier attack: motivation

- The model misclassifies if there is at least one index where the model's prediction is greater than the prediction of the correct index:

$$\max_{i \neq c} f_i(x) - f_c(x) > 0$$

- Thus we can rewrite (1):

$$\begin{aligned} &\underset{\delta}{\text{minimize}} \quad \|\delta\| \\ &\text{subject to} \quad \max_{i\neq c} f_i(x+\delta) - f_c(x+\delta) > 0. \end{aligned} \tag{2}$$

- The barrier method is a standard tool in optimization for solving problems such as (2) with inequality constraints. Inequality constraints are incorporated into the objective function via a penalty term, which is infinite if a constraint is violated. If a constraint is far from being active, then the penalty term should be small.
- The negative logarithm is an ideal choice:

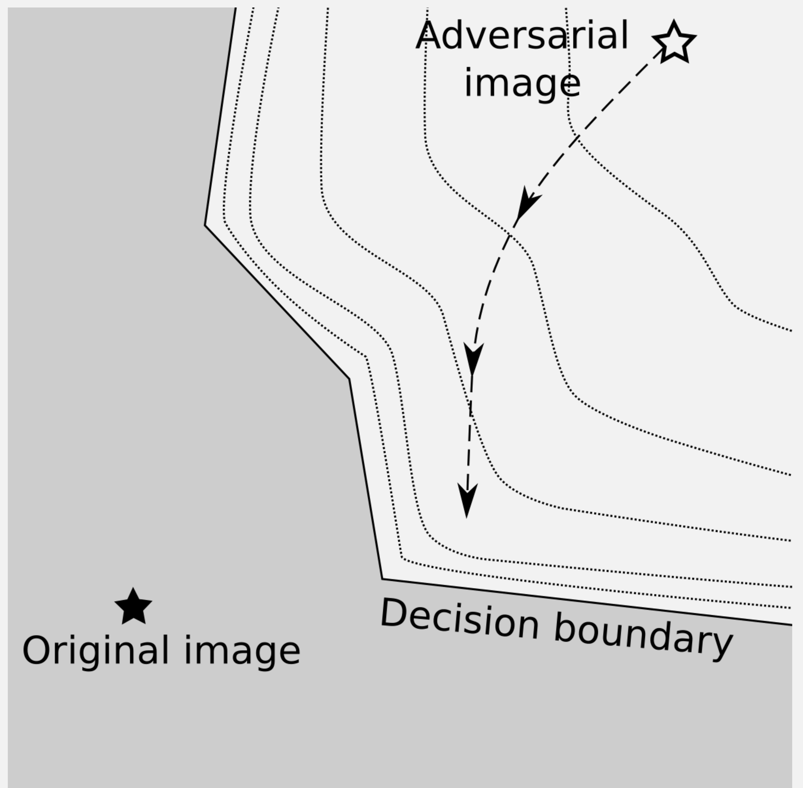
$$\min_{\delta} \|\delta\| - \lambda \log\left(f_{\max}(x+\delta) - f_c(x)\right) \tag{3}$$

Examples

ℓ_∞ attacks, MNIST

ℓ_∞ attacks, CIFAR10

Algorithmic outline

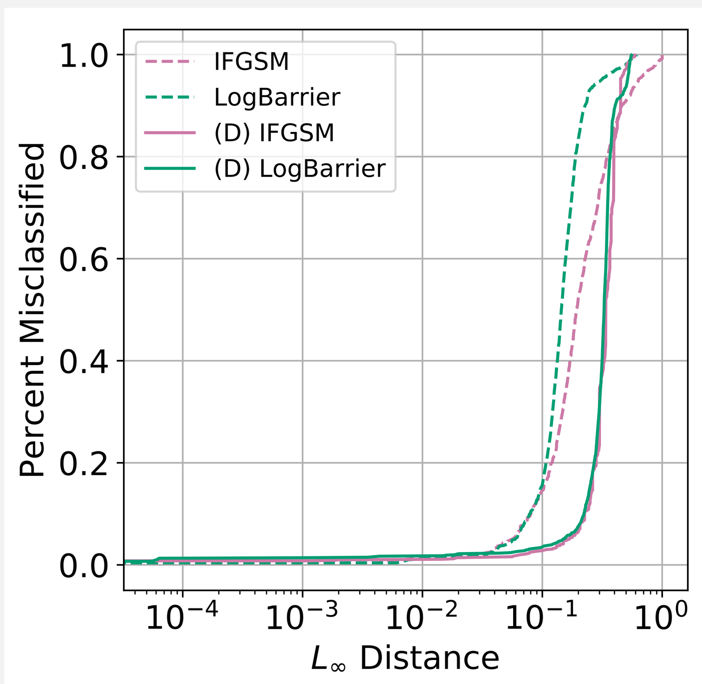


Rather than solve (1), the LogBarrier attack aims for an approximate solution through (3), as follows.

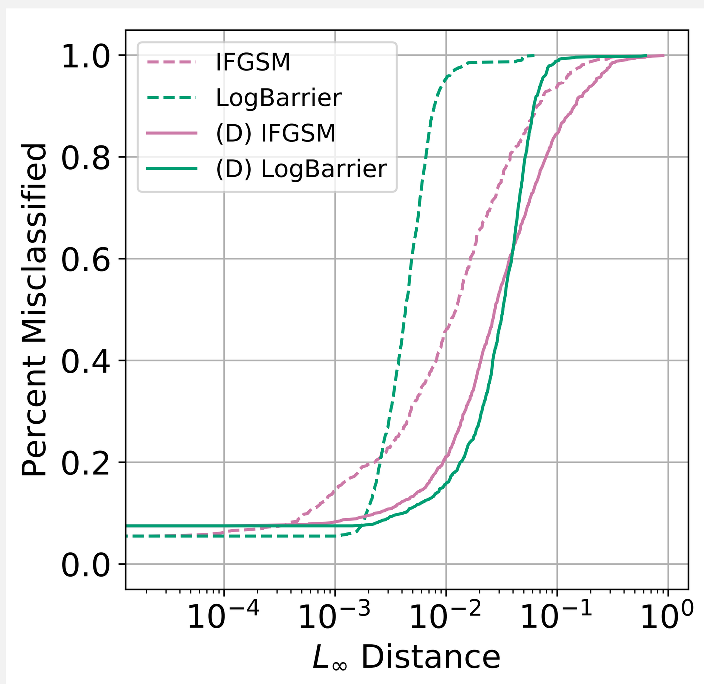
1. Initialize with a mis-classified image, *far from the original image*. For example, this could be done by adding random noise to the original image, until it is misclassified.
2. Fix λ and solve (3) via gradient descent.
3. Continue repeating step 2 while decreasing λ until a desired solution tolerance is achieved.

As λ decreases, solutions move closer to the decision boundary (see above Figure).

Attack curves



MNIST



CIFAR10

Attack curves measured in ℓ_∞, on MNIST and CIFAR10 networks. Two types of networks are compared: an undefended network, and a defended network (denoted (D)), trained using the same architecture as the undefended network with adversarial training. The LogBarrier attack requires a smaller adversarial distance to attack all images, compared to IFGSM.

Results & Discussion

- The LogBarrier attack achieves similar or better attack distances than current state-of-the-art attacks on standard datasets
- The LogBarrier attack performs significantly better on challenging images (those that require large perturbations for misclassification)
- The LogBarrier attack performs well on adversarially defended models (through adversarial training [3]): the distance needed to perturb *all* images is significantly smaller than other attacks.
- Although the LogBarrier attack uses gradients, we show it overcomes gradient obfuscation, a common pitfall of other gradient-based attacks

Comparison of attacks at specified perturbation size

Table: Percent misclassification of the networks at a specified perturbation size, for attacks measured in ℓ_2. Because we are measuring the strength of adversarial attacks, at a given adversarial distance, a higher percentage misclassified is better.

	MNIST	CIFAR10	Imagenet-1K
$\|\delta\|_2$	2.3	120/255	1
LogBarrier	**99.10**	**99.90**	**98.40**
Carlini-Wagner [4]	98.50	90.40	74.86
PGD	52.58	59.80	90.00
Boundary Attack [5]	97.20	99.60	48.80

Table: Percent misclassification of the networks at a specified perturbation size, for attacks measured in ℓ_∞. Higher percentage misclassified is better.

	MNIST	CIFAR10	Imagenet-1K
$\|\delta\|_\infty$	0.3	8 / 255	8 / 255
LogBarrier	**94.80**	**98.70**	95.20
IFGSM	73.40	75.80	**99.60**

References

[1] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *CoRR*, abs/1312.6199, 2013.
[2] Jorge Nocedal and Stephen Wright. *Numerical optimization*. Springer Science & Business Media, 2006.
[3] Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *CoRR*, abs/1706.06083, 2017.
[4] Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.
[5] Wieland Brendel, Jonas Rauber, and Matthias Bethge. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. In *International Conference on Learning Representations*, 2018.