Towards Evaluating the Robustness of Neural Networks

Nicholas Carlini and David Wagner
University of California, Berkeley

Background: Adversarial Examples

For a classification neural network $F(x)$

Given an input X classified as label L ...

... it is easy to find an X' close to X

... so that $F(X') \neq L$

Motivation:
Why should we care?

Distance Metrics

"Adversarial examples are close to the original"

How do we define **close**?

This is what lets us compare attacks.

In what domain? Images.

Distance Metrics

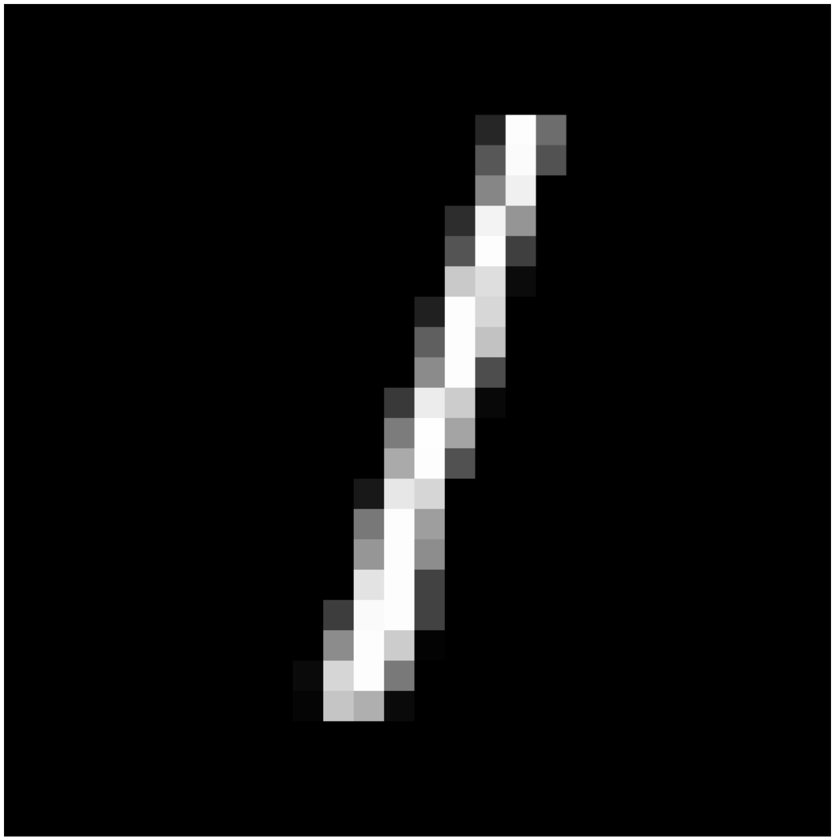
L_p distance metrics:

L_0 - number of pixels changed

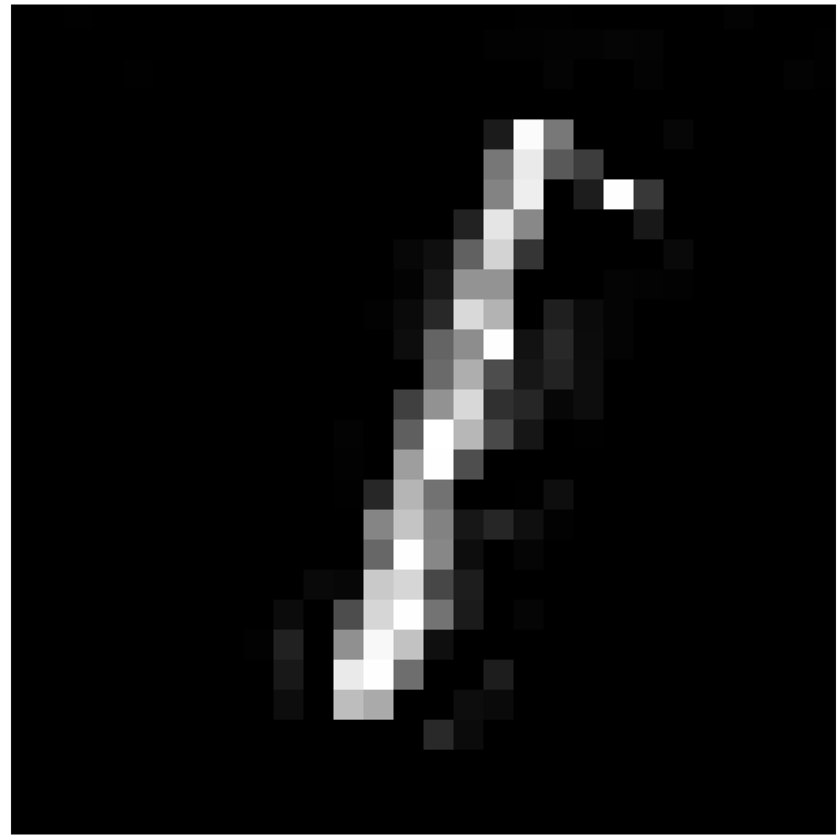
L_2 - standard Euclidian distance

$L_{infinity}$ - amount each pixel can be changed

If any L_p distance is small, the two images should be visually similar



Classified as a 1



Classified as a 0

For this talk:

Assume complete knowledge of model parameters

(but lots of work exists for other threat models)

Two ways to evaluate robustness:

1. Construct a proof of robustness
2. Demonstrate constructive attack

Proving Robustness

It is possible to prove robustness

... for specific input points

... on simple datasets (e.g., MNIST)

... for small networks (e.g., 100 neurons)

... for ReLU activations

N Carlini, G Kat, C Barrett, and D Dill. "Provably Minimally-Distorted Adversarial Examples." Under Submission to ICML.

Finding Adversarial Examples

Formulation: given input x, find x′ where
minimize $d(x,x')$
such that $F(x') = T$
x′ is "valid"

Gradient Descent to the rescue?

Non-linear constraints are hard

Reformulation

Formulation:

minimize $d(x,x') + g(x')$

such that x' is "valid"

Where $g(x')$ is some kind of loss function on how close $F(x')$ is to target T

$g(x')$ is small if $F(x') = T$

$g(x')$ is large if $F(x') != T$

Reformulation

For example

$$g(x') = (1\text{-}F(x')_T)$$

If F(x′) says the probability of T is 1:

$$g(x') = (1\text{-}F(x')_T) = (1\text{-}1) = 0$$

F(x′) says the probability of T is 0:

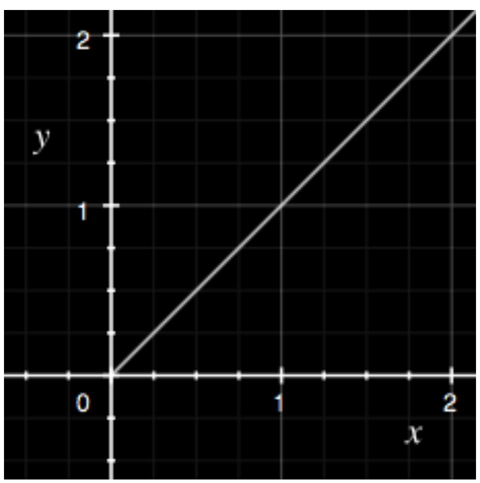
$$g(x') = (1\text{-}F(x')_T) = (1\text{-}0) = 1$$

Does this work?

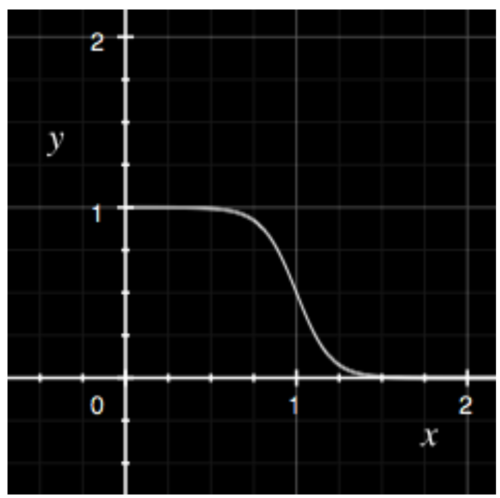
Formulation:
minimize d(x,x′) + g(x′)
such that x′ is "valid"

Problem 1:
Global minimum is not an adversarial example

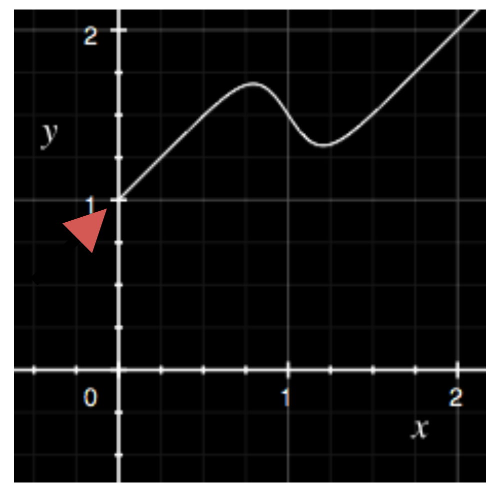
d(x,x′) + g(x′)



+



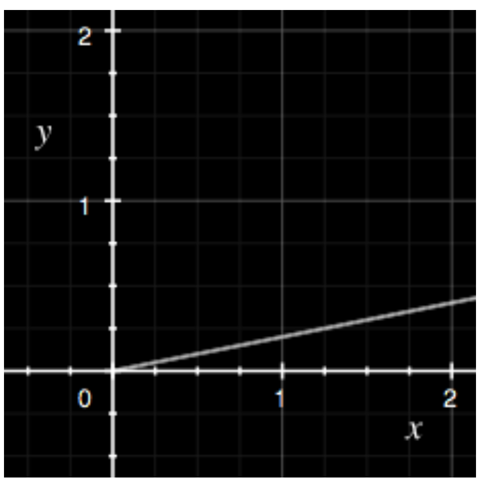
=



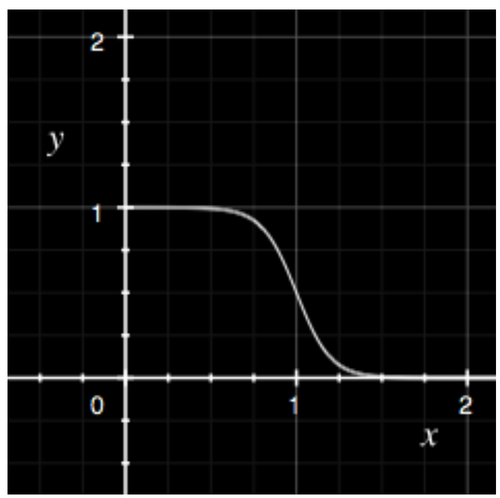
Does this work?

Formulation:
minimize $d(x,x')/5 + g(x')$
such that x' is "valid"

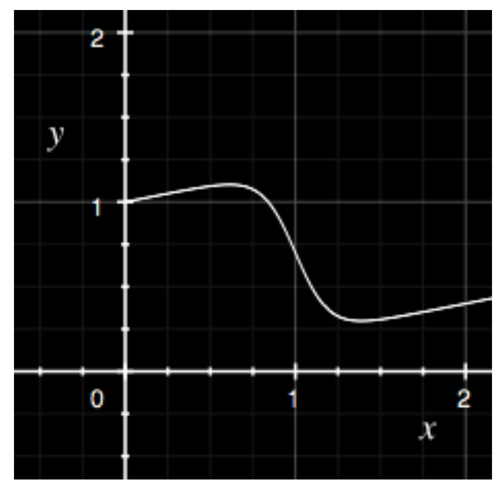
$d(x,x')/5$ + $g(x')$



+



=



Does this work?

Formulation:

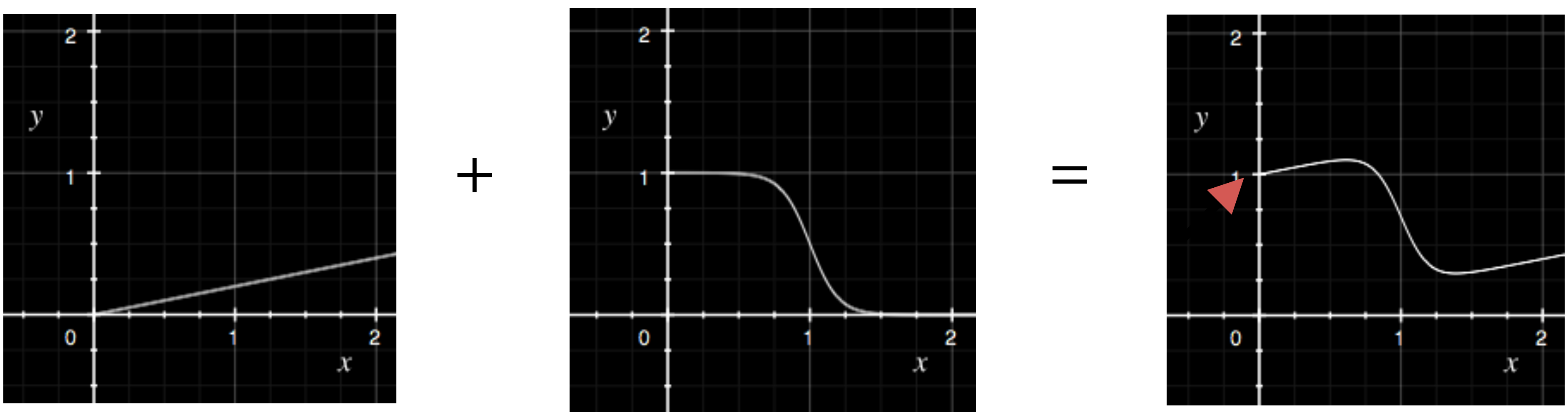
minimize $\quad d(x,x')/5 + g(x')$

such that $\quad x'$ is "valid"

Problem 2:

Gradient direction does not point toward the global minimum

$d(x,x')/5 \quad + \quad g(x')$



Does this work?

Formulation:

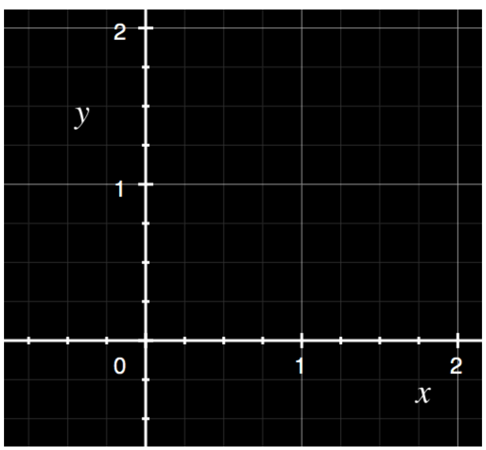
minimize $d(x,x')/1e10 + g(x')$

such that x' is "valid"

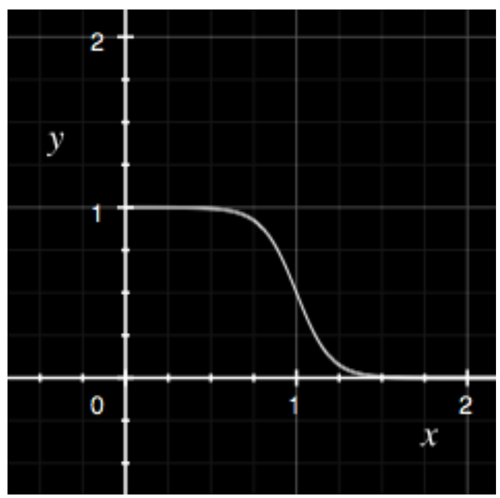
Problem 3:

Global minimum is not the minimally perturbed adversarial example

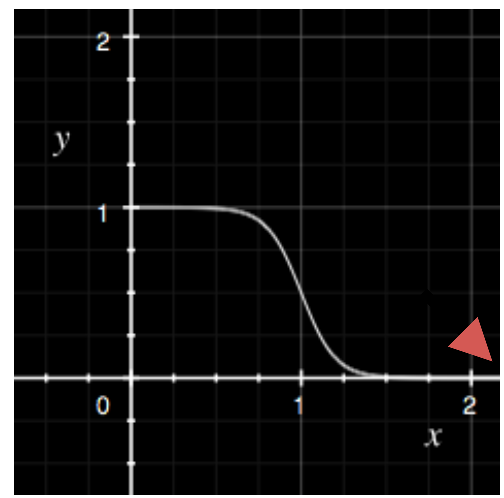
$d(x,x')/1e10$ + $g(x')$



+



=



Constructing a better loss function

Global minimum at the decision boundary

Gradient points towards the global minimum

$$\max\left(\max_{t'\neq t}\left\{\log(F(x)'_t)\right\} - \log(F(x)_t), 0\right)$$

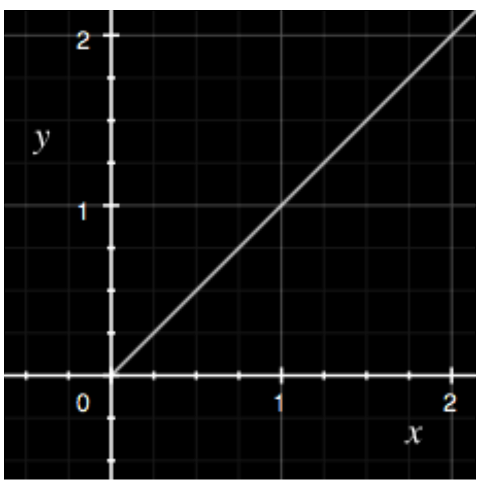
Improved Formulation

Formulation:

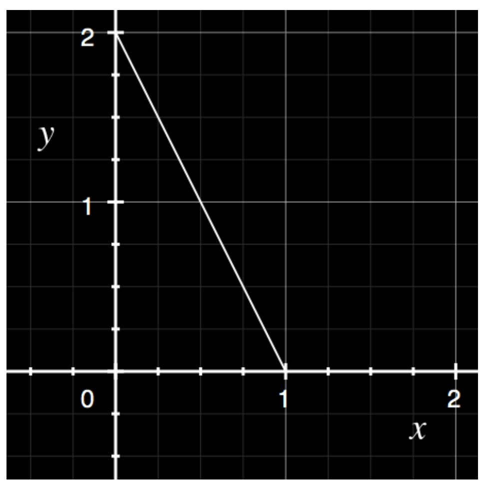
minimize $d(x,x') + g(x')$

such that x' is "valid"

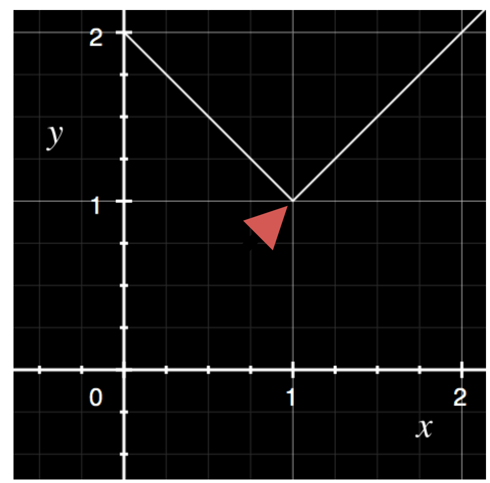
$d(x,x')$ + $g(x')$

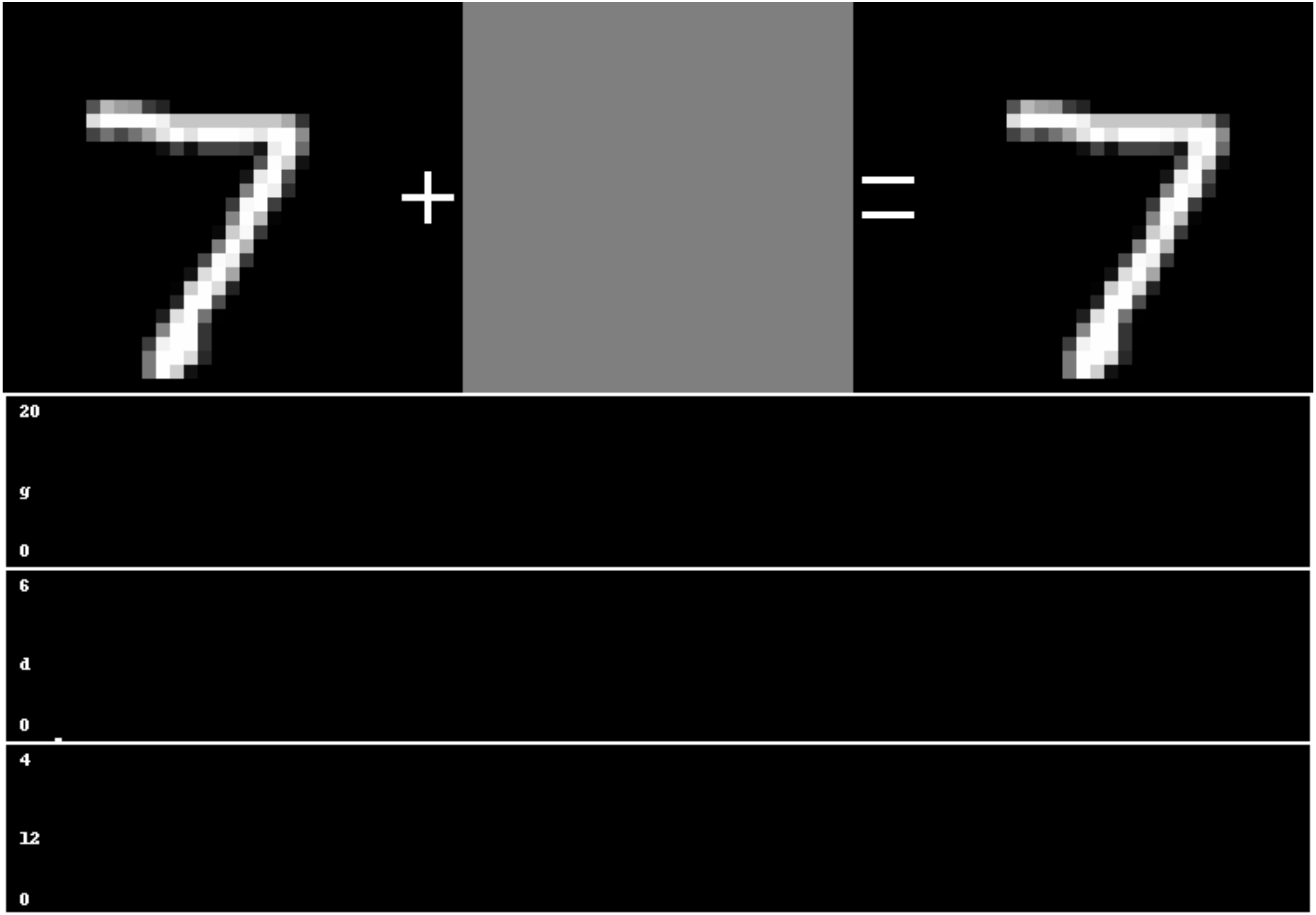


+



=




+
=
20
g
0
6
d
0
4
12
0

L_0 from L_2

First attempt:

minimize $\quad d(x,x') + g(x')$
such that $\quad x'$ is "valid"

Where the distance d is the L_0 distance

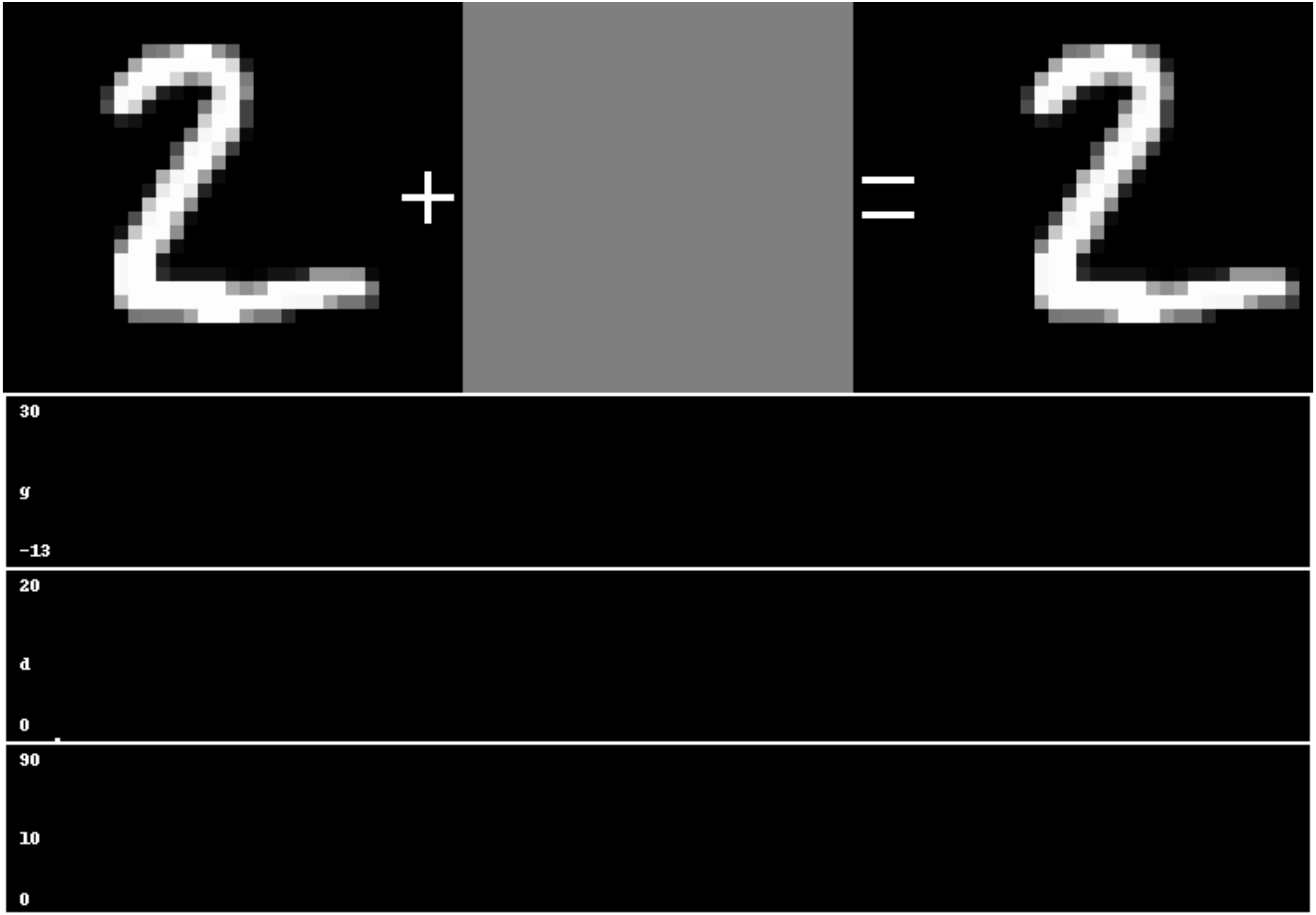
L_0 from L_2

Solve the L_2 minimization problem and identify the least changed pixel

Force that pixel to remain constant

Re-solve the L_2 minimization problem with that pixel fixed at the initial value

Repeat, finding the new least-changed pixel



30
g
−13
20
d
0
90
10
0

$L_{infinity}$ from L_2

Formulation:

minimize   d(x,x') + g(x')

such that   x is "valid"

$L_{infinity}$ from L_2

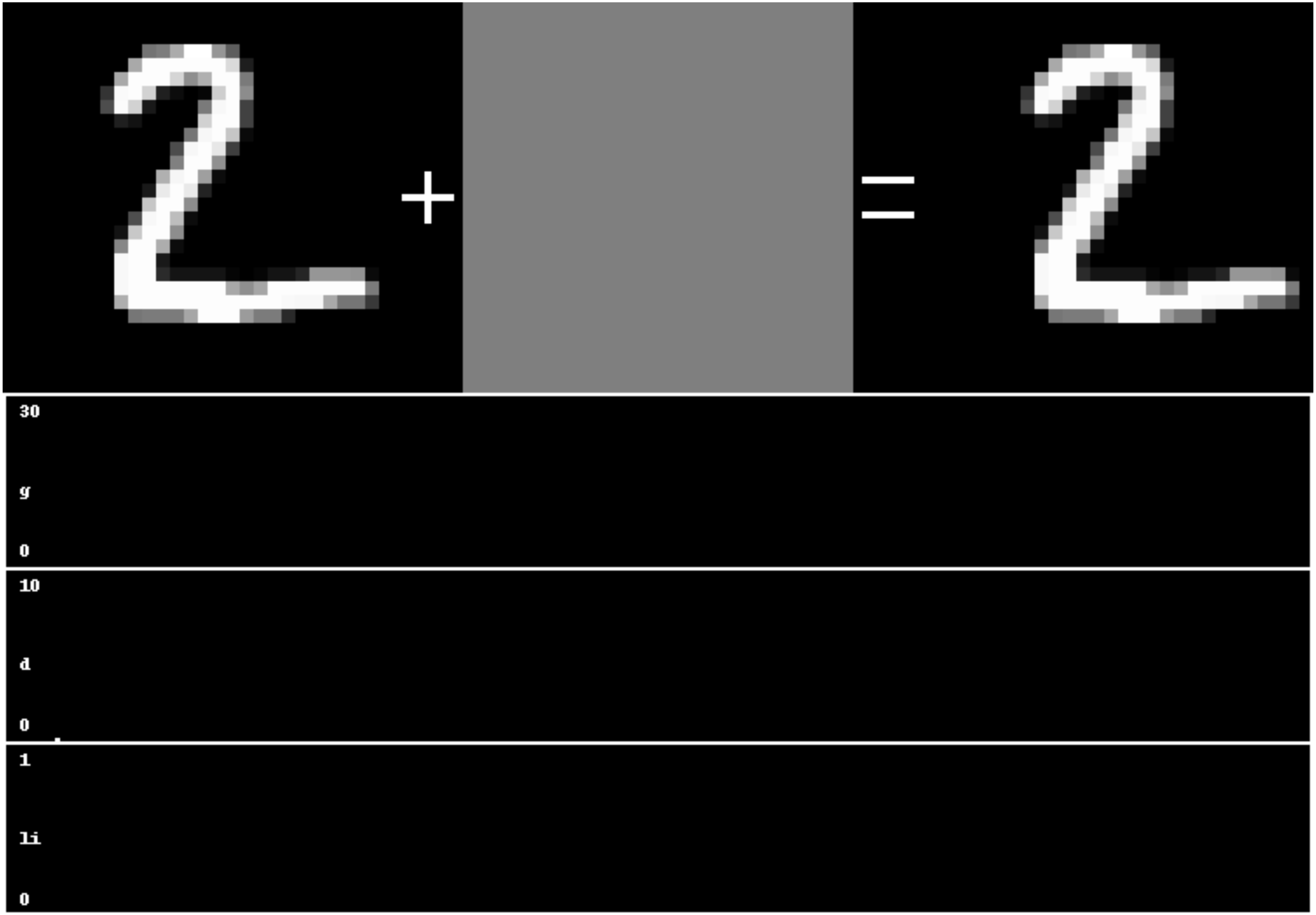
Initially set a budget $\Delta=1$

Formulation:

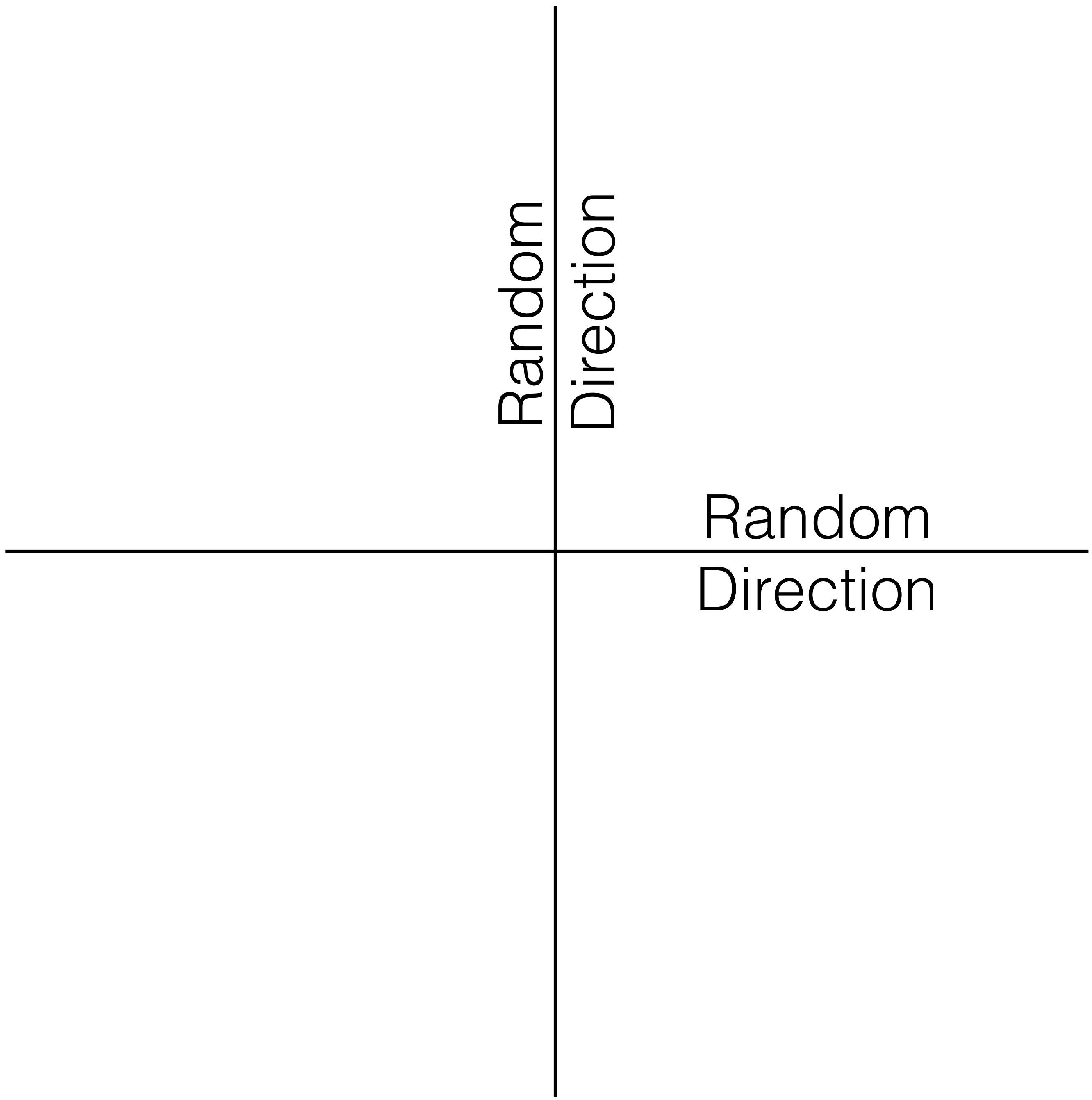
minimize $\quad \text{sum}[\max(|x_i - x'_i| - \Delta, 0)] + g(x')$

such that $\quad$ x is "valid"

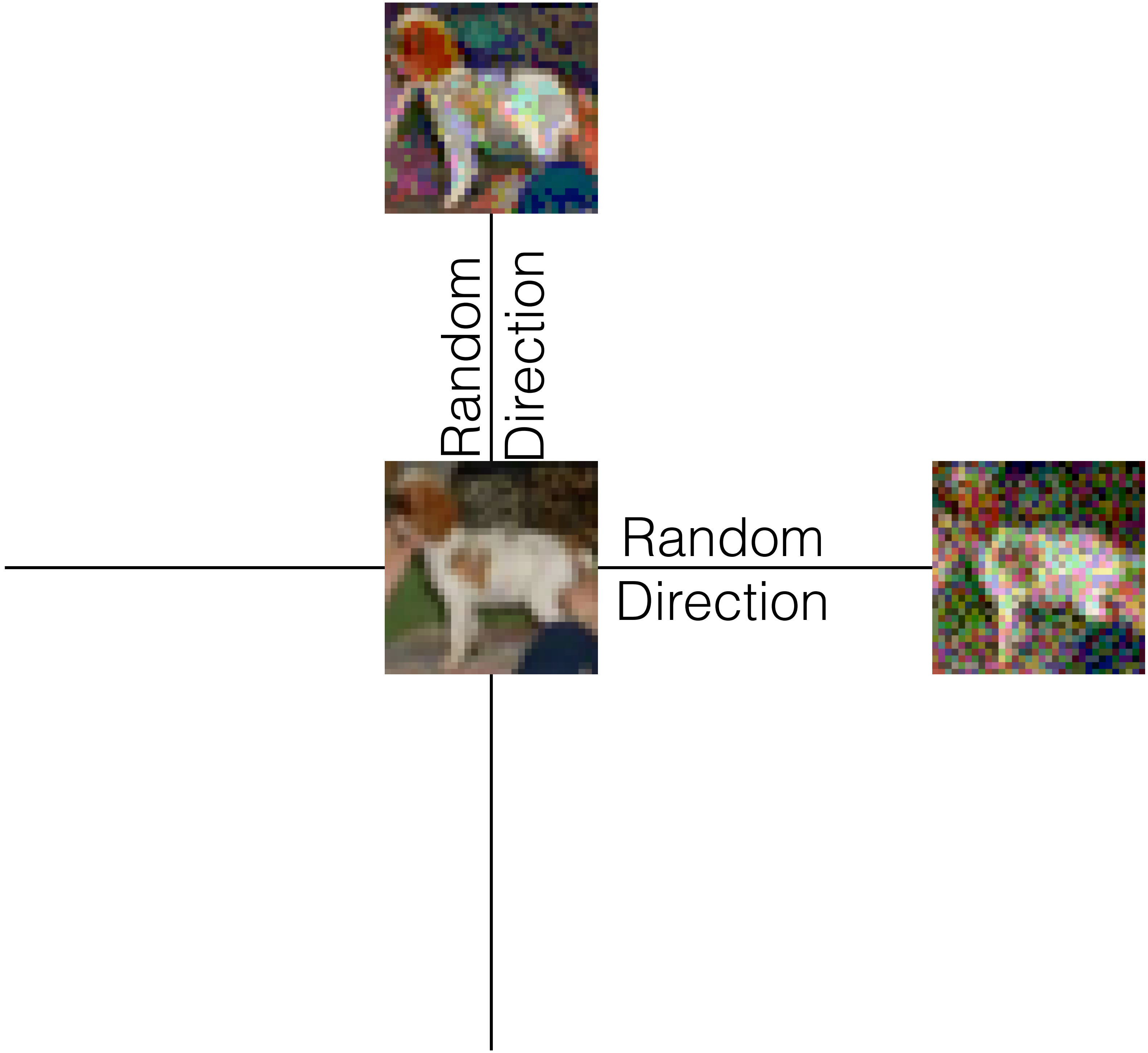
Decrease Δ and solve again

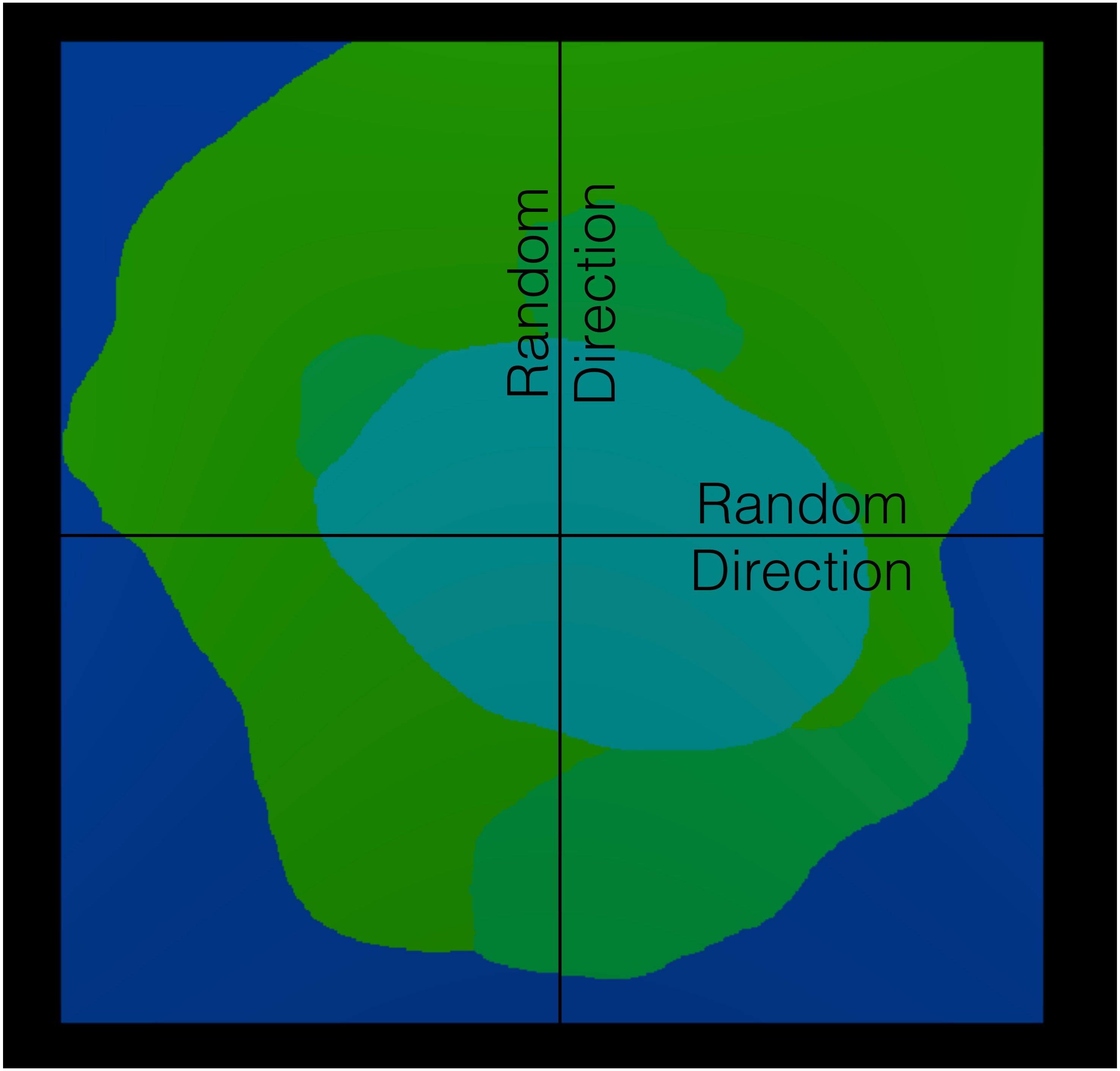

+
=
30
g
0
10
d
0
1
li
0

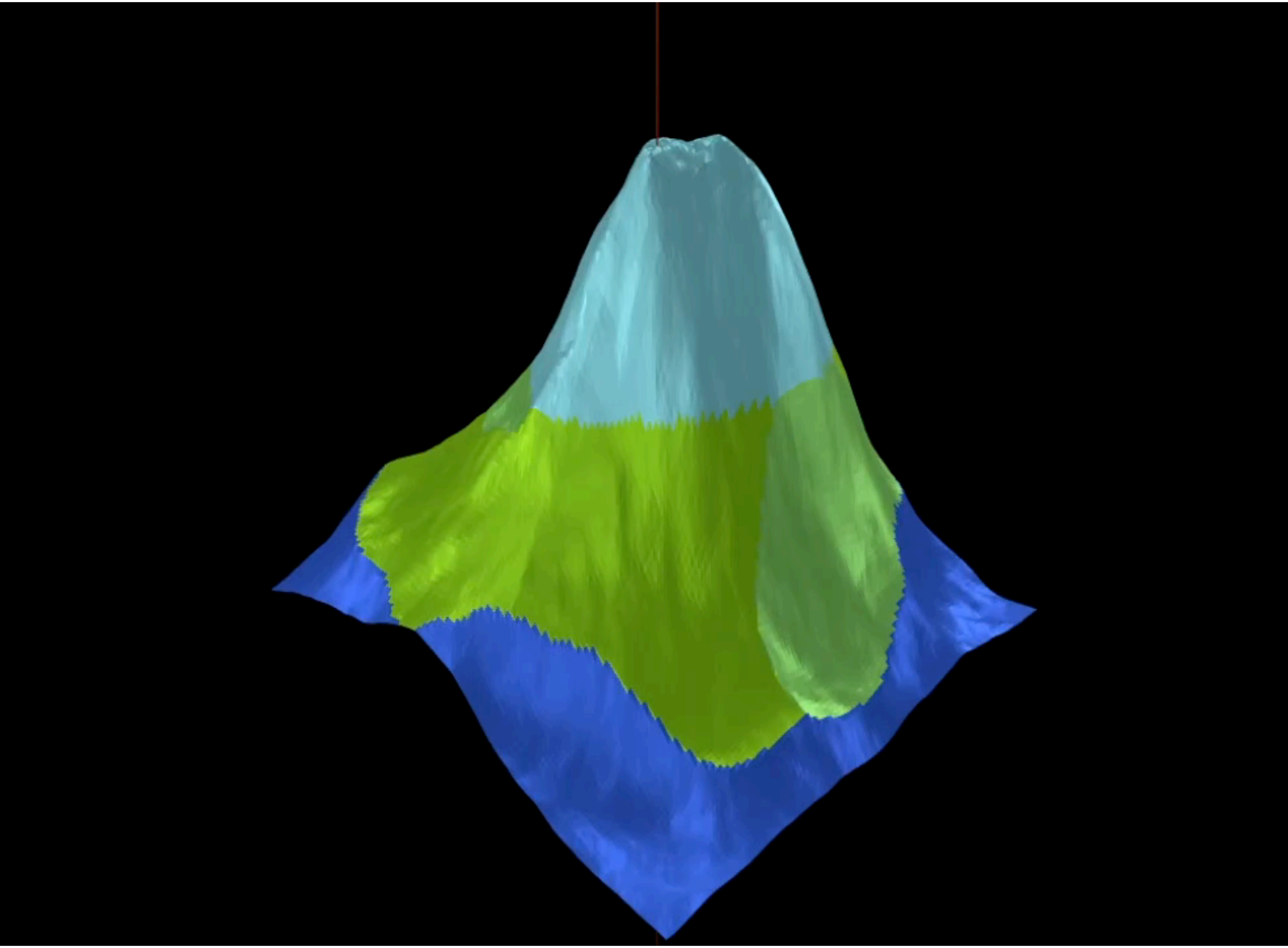
Visualizations

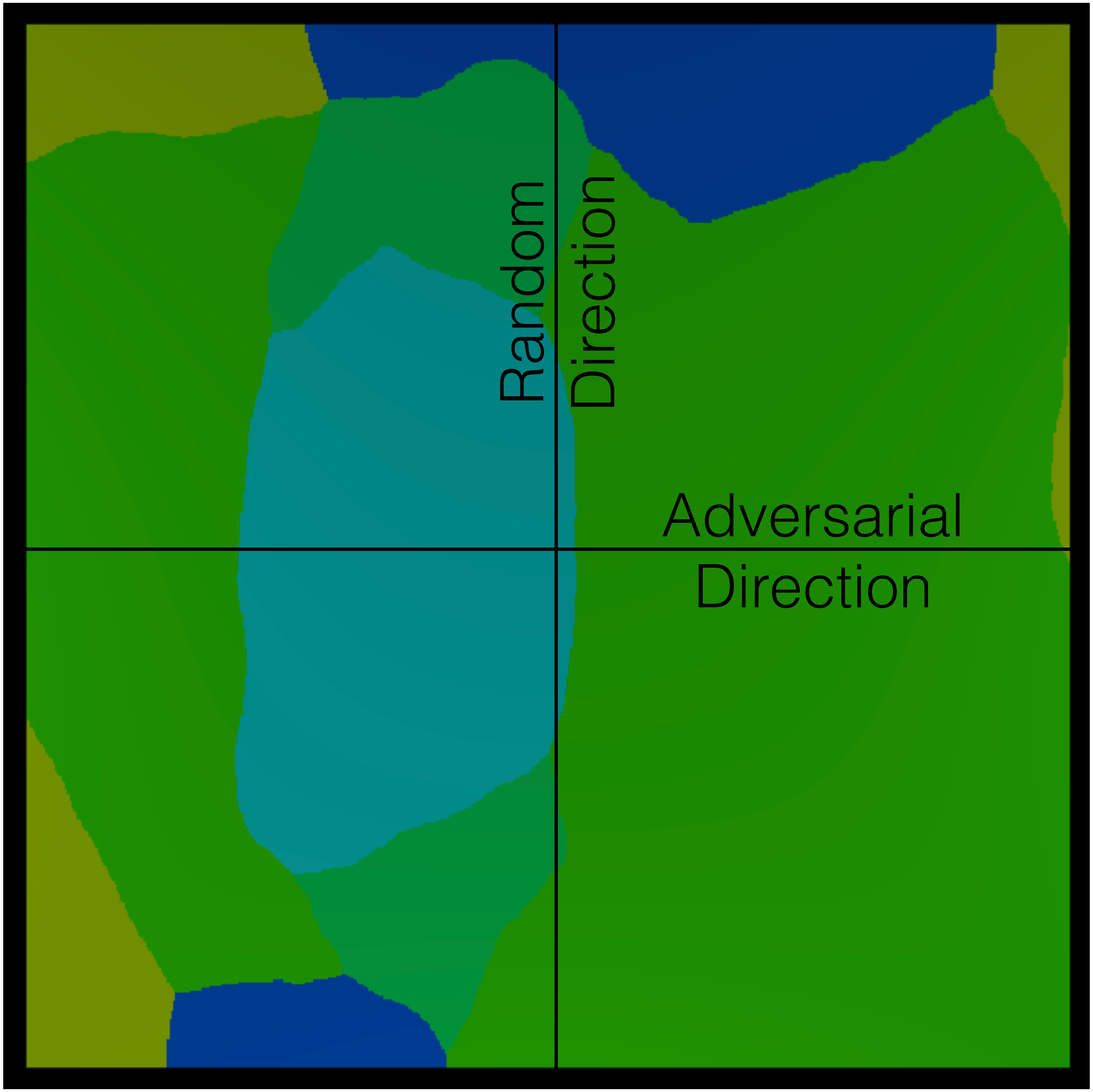


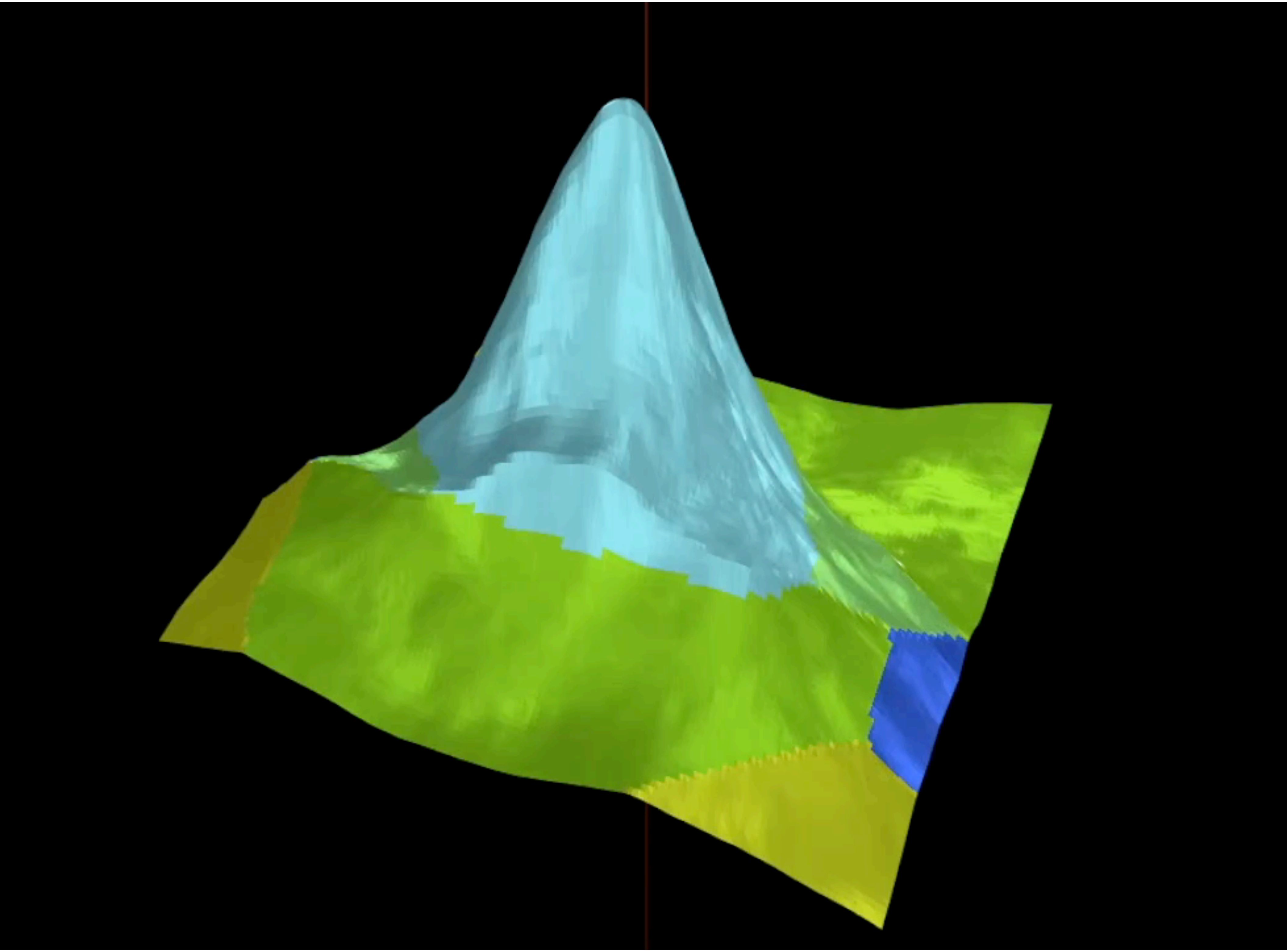
Random Direction
Random Direction


Random Direction
Random Direction


Random Direction
Random Direction




Random Direction
Adversarial Direction



Is this attack useful?

This attack breaks almost everything

N Carlini and D Wagner, "Defensive Distillation is Not Robust to Adversarial Examples". 2016

N Carlini and D Wagner. "Adversarial Examples are not Easily Detected". AISEC. 2017

N Carlini and D Wagner. "MagNet and "Efficient Defenses against Adversarial Attack" are Not Robust to Adversarial Examples". 2017

A Athalye, N Carlini and D Wagner. "Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples". Under submission to ICML.

	Best Case						Average Case						Worst Case					
	Change of Variable		Clipped Descent		Projected Descent		Change of Variable		Clipped Descent		Projected Descent		Change of Variable		Clipped Descent		Projected Descent	
	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob
f_1	2.46	100%	2.93	100%	2.31	100%	4.35	100%	5.21	100%	4.11	100%	7.76	100%	9.48	100%	7.37	100%
f_2	4.55	80%	3.97	83%	3.49	83%	3.22	44%	8.99	63%	15.06	74%	2.93	18%	10.22	40%	18.90	53%
f_3	4.54	77%	4.07	81%	3.76	82%	3.47	44%	9.55	63%	15.84	74%	3.09	17%	11.91	41%	24.01	59%
f_4	5.01	86%	6.52	100%	7.53	100%	4.03	55%	7.49	71%	7.60	71%	3.55	24%	4.25	35%	4.10	35%
f_5	1.97	100%	2.20	100%	1.94	100%	3.58	100%	4.20	100%	3.47	100%	6.42	100%	7.86	100%	6.12	100%
f_6	1.94	100%	2.18	100%	1.95	100%	3.47	100%	4.11	100%	3.41	100%	6.03	100%	7.50	100%	5.89	100%
f_7	1.96	100%	2.21	100%	1.94	100%	3.53	100%	4.14	100%	3.43	100%	6.20	100%	7.57	100%	5.94	100%

TABLE III

EVALUATION OF ALL COMBINATIONS OF ONE OF THE SEVEN POSSIBLE OBJECTIVE FUNCTIONS WITH ONE OF THE THREE BOX CONSTRAINT ENCODINGS. WE SHOW THE AVERAGE L_2 DISTORTION, THE STANDARD DEVIATION, AND THE SUCCESS PROBABILITY (FRACTION OF INSTANCES FOR WHICH AN ADVERSARIAL EXAMPLE CAN BE FOUND). EVALUATED ON 1000 RANDOM INSTANCES. WHEN THE SUCCESS IS NOT 100%, MEAN IS FOR SUCCESSFUL ATTACKS ONLY.

	Best Case				Average Case				Worst Case			
	MNIST		CIFAR		MNIST		CIFAR		MNIST		CIFAR	
	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob
Our L_0	10	100%	7.4	100%	19	100%	15	100%	36	100%	29	100%
Our L_2	1.7	100%	0.36	100%	2.2	100%	0.60	100%	2.9	100%	0.92	100%
Our L_∞	0.14	100%	0.002	100%	0.18	100%	0.023	100%	0.25	100%	0.038	100%

TABLE VI
COMPARISON OF OUR ATTACKS WHEN APPLIED TO DEFENSIVELY DISTILLED NETWORKS. COMPARE TO TABLE IV FOR UNDISTILLED NETWORKS.

	Best Case				Average Case				Worst Case			
	MNIST		CIFAR		MNIST		CIFAR		MNIST		CIFAR	
	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob	mean	prob
Our L_0	8.5	100%	5.9	100%	16	100%	13	100%	33	100%	24	100%
JSMA-Z	20	100%	20	100%	56	100%	58	100%	180	98%	150	100%
JSMA-F	17	100%	25	100%	45	100%	110	100%	100	100%	240	100%
Our L_2	1.36	100%	0.17	100%	1.76	100%	0.33	100%	2.60	100%	0.51	100%
Deepfool	2.11	100%	0.85	100%	—	-	—	-	—	-	—	-
Our L_∞	0.13	100%	0.0092	100%	0.16	100%	0.013	100%	0.23	100%	0.019	100%
Fast Gradient Sign	0.22	100%	0.015	99%	0.26	42%	0.029	51%	—	0%	0.34	1%
Iterative Gradient Sign	0.14	100%	0.0078	100%	0.19	100%	0.014	100%	0.26	100%	0.023	100%

TABLE IV
COMPARISON OF THE THREE VARIANTS OF TARGETED ATTACK TO PREVIOUS WORK FOR OUR MNIST AND CIFAR MODELS. WHEN SUCCESS RATE IS NOT 100%, THE MEAN IS ONLY OVER SUCCESSES.

	Untargeted		Average Case		Least Likely	
	mean	prob	mean	prob	mean	prob
Our L_0	48	100%	410	100%	5200	100%
JSMA-Z	-	0%	-	0%	-	0%
JSMA-F	-	0%	-	0%	-	0%
Our L_2	0.32	100%	0.96	100%	2.22	100%
Deepfool	0.91	100%	-	-	-	-
Our L_∞	0.004	100%	0.006	100%	0.01	100%
FGS	0.004	100%	0.064	2%	-	0%
IGS	0.004	100%	0.01	99%	0.03	98%

TABLE V

COMPARISON OF THE THREE VARIANTS OF TARGETED ATTACK TO PREVIOUS WORK FOR THE INCEPTION V3 MODEL ON IMAGENET. WHEN SUCCESS RATE IS NOT 100%, THE MEAN IS ONLY OVER SUCCESSES.

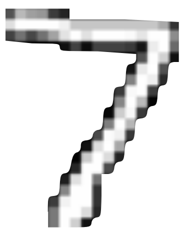
Case studies on evaluating defenses to adversarial examples

Defense Idea #1:

Additional Neural Network Detection

Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischo. 2017. On Detecting Adversarial Perturbations. In International Conference on Learning Representations.

Normal Classifier



Classifier

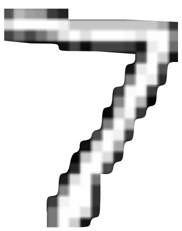
7

Normal Classifier

Classifier

8

Detector & Classifier



Detector

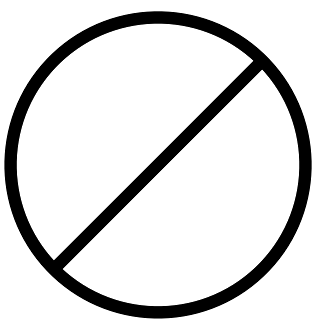
Classifier

7

Detector & Classifier

Detector

Classifier



Training an adversarial example detector

Normal Training

( , 7)

( , 3)

Training

F

Detection Training (1)

(, 7)

(, 3)

(, n)

(, n)

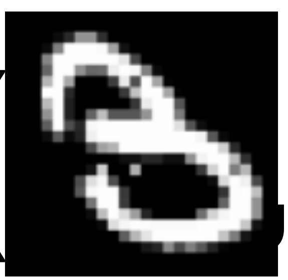
Attack

Detection Training (2)

(, y)

(, y)

(, n)

(, n)

Training

G

Sounds great.

Sounds great.

But we already know it's easy to fool neural networks ...

... so just construct adversarial examples to

1. be misclassified
2. not be detected

Breaking Detection Adversarial Training

minimize $\quad d(x,x') + g(x')$
such that $\quad x'$ is "valid"

Old: $g(x')$ measures loss of **classifier** on x'

Breaking Detection Adversarial Training

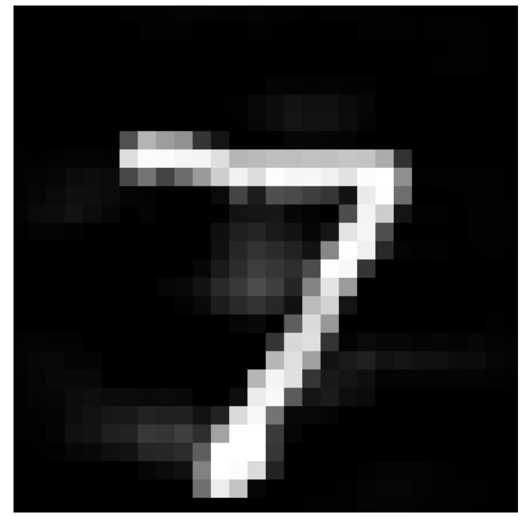
minimize	$d(x,x') + g(x') + h(x')$
such that	x' is "valid"

Old: $g(x')$ measures loss of **classifier** on x'

New: $h(x')$ measures loss of **detector** on x'

Original

Adversarial (unsecured)



Adversarial (with detector)

Defense Idea #2:

Thermometer Encoding

Jacob Buckman, Aurko Roy, Colin Raffel, and Ian Goodfellow. 2018. Thermometer encoding: One hot way to resist adversarial examples. In International Conference on Learning Representations.

Problem:

Neural Networks are "overly linear"

Thermometer Encoding

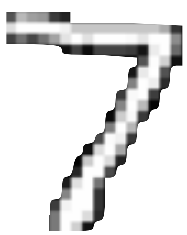
Break linearity by changing input representation

T(0.13) = 1 1 0 0 0 0 0 0 0 0

T(0.66) = 1 1 1 1 1 1 0 0 0 0

T(0.97) = 1 1 1 1 1 1 1 1 1 1

Standard Neural Network



F

0.000
0.102
0.001
0.002
0.001
0.001
0.001
0.890
0.000
0.002

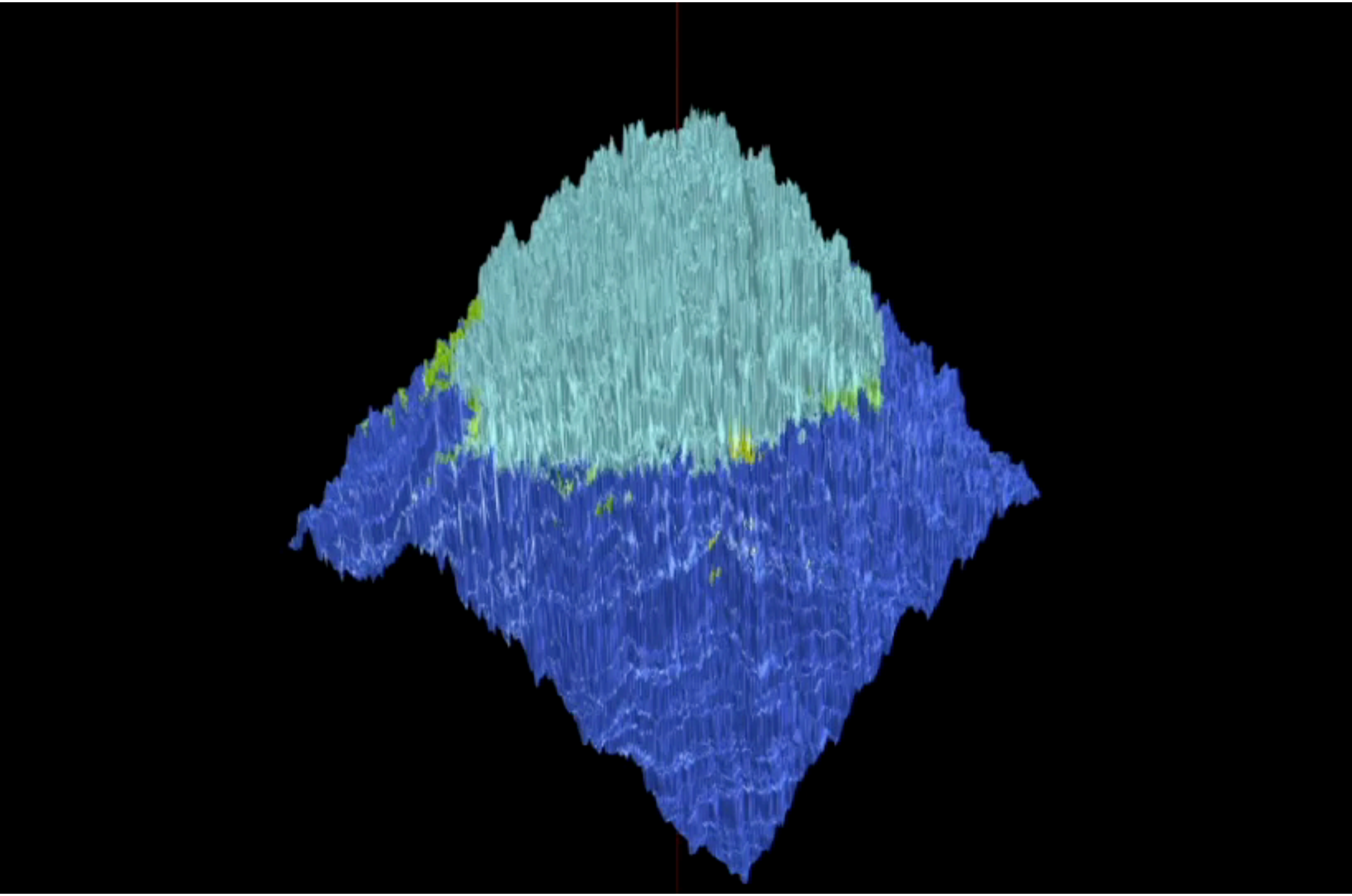
With Thermometer Encoding

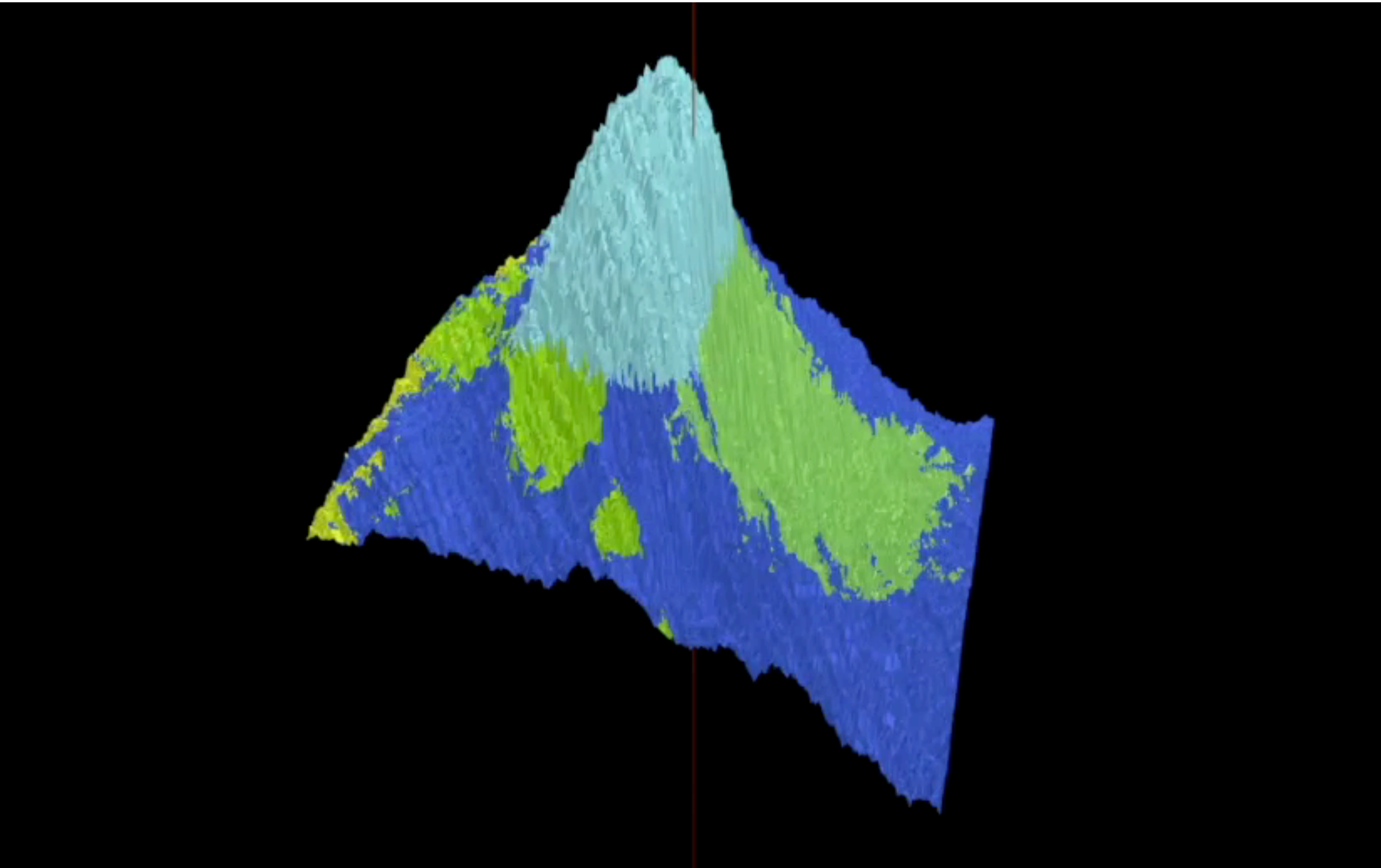
Claims:

On CIFAR,
with distortion 8/255,
accuracy of 50%

(compared to 0%)

Unfortunately, thermometer encoding only causes gradient descent to fail





Defense Idea #3:

Adversarial Retraining

A Madry, A Makelov, L Schmidt, D Tsipras, and A Vladu. Towards deep learning models resistant to adversarial attacks. 2018. International Conference on Learning Representations.

Adversarial Training

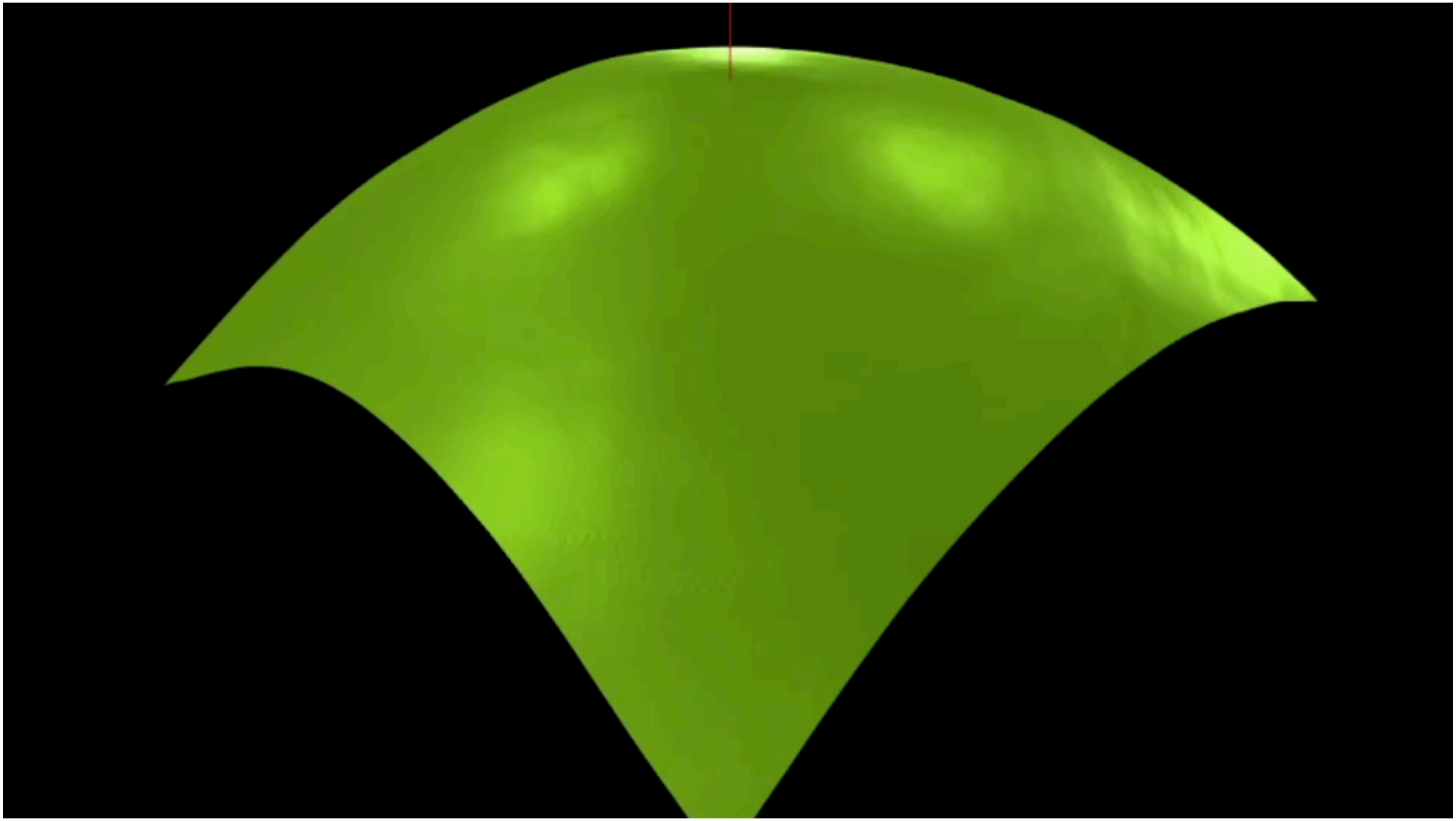
Given training data (X,Y)

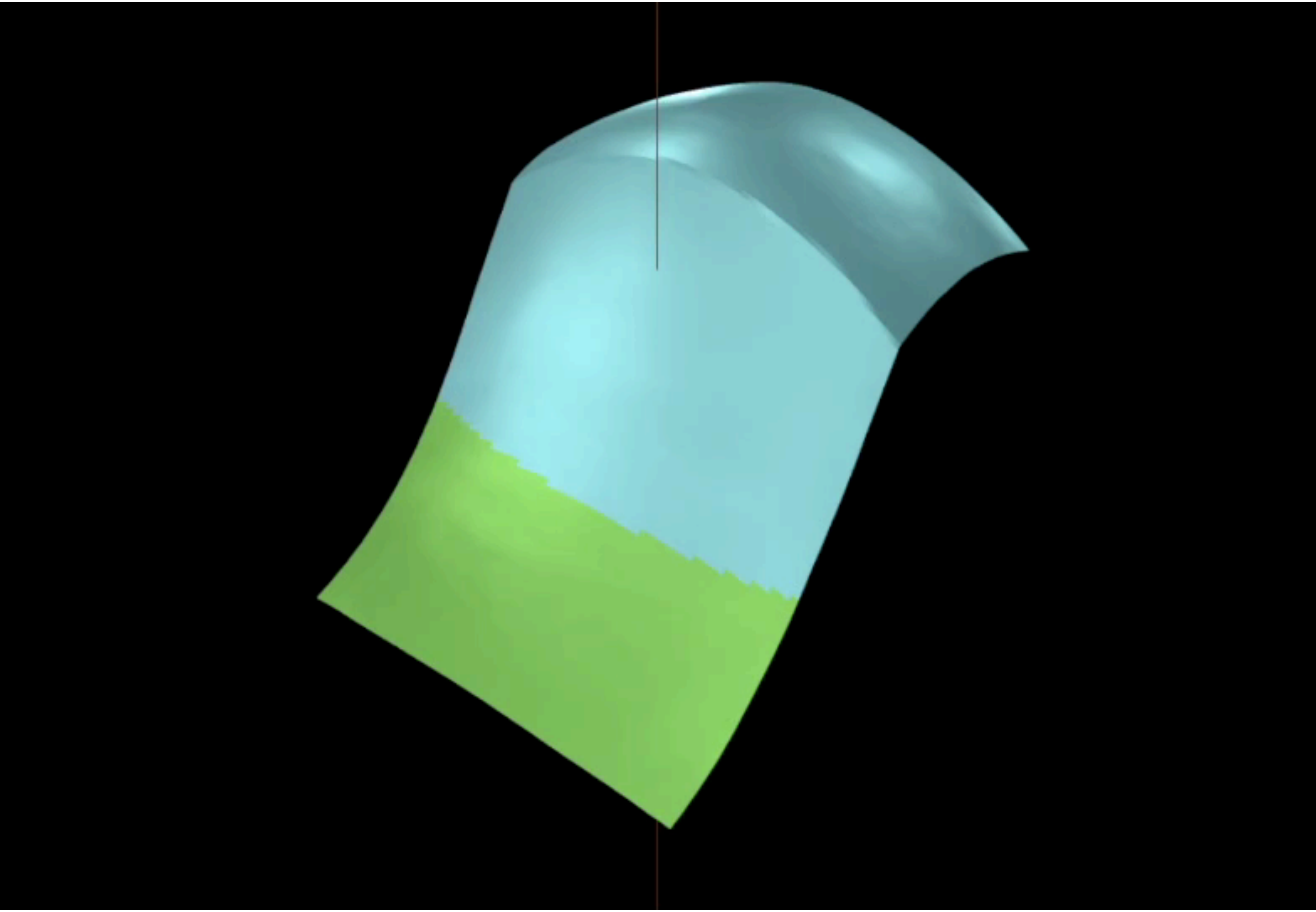
Sample a minibatch (x,y)

Generate the adversarial minibatch (x',y)

Train on (x',y)

Repeat until convergence





... so that's images

what about other domains?

Audio has these same issues, too

N Carlini and D Wagner. "Audio Adversarial Examples: Targeted Attacks on Speech-to-Text". 2018.

"now I would drift gently off to dream land"

[adversarial]

It was the best of times, it was the worst of times, it was the age of wisdom, it was the age of foolishness, it was the epoch of belief, it was the epoch of incredulity original or adversarial?

original or adversarial?

On audio, traditional ML methods are not vulnerable to adversarial examples

Questions?

Nicholas Carlini
https://nicholas.carlini.com
npc@berkeley.edu