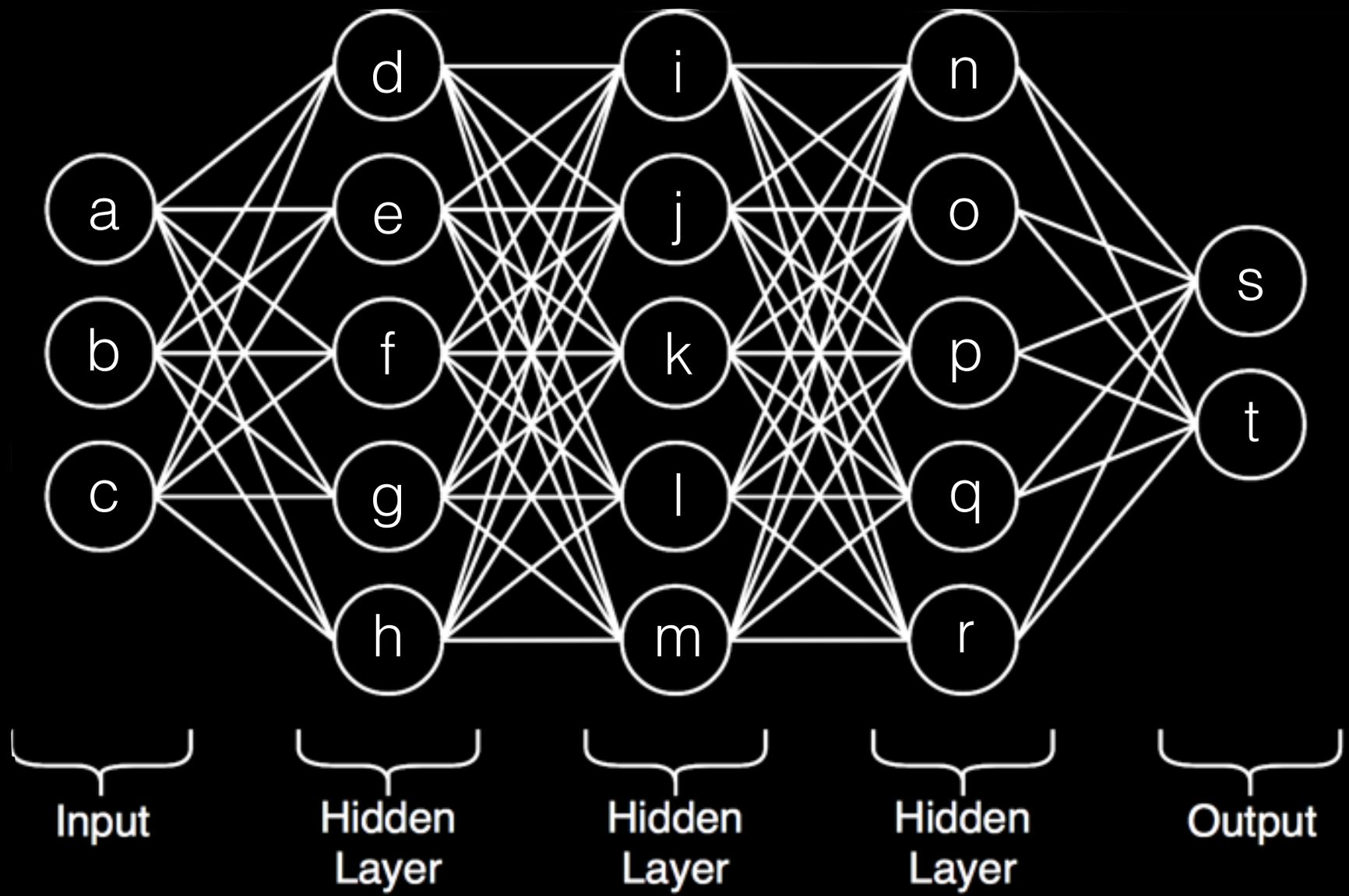
Towards Evaluating the Robustness of Neural Networks

Nicholas Carlini and David Wagner
University of California, Berkeley

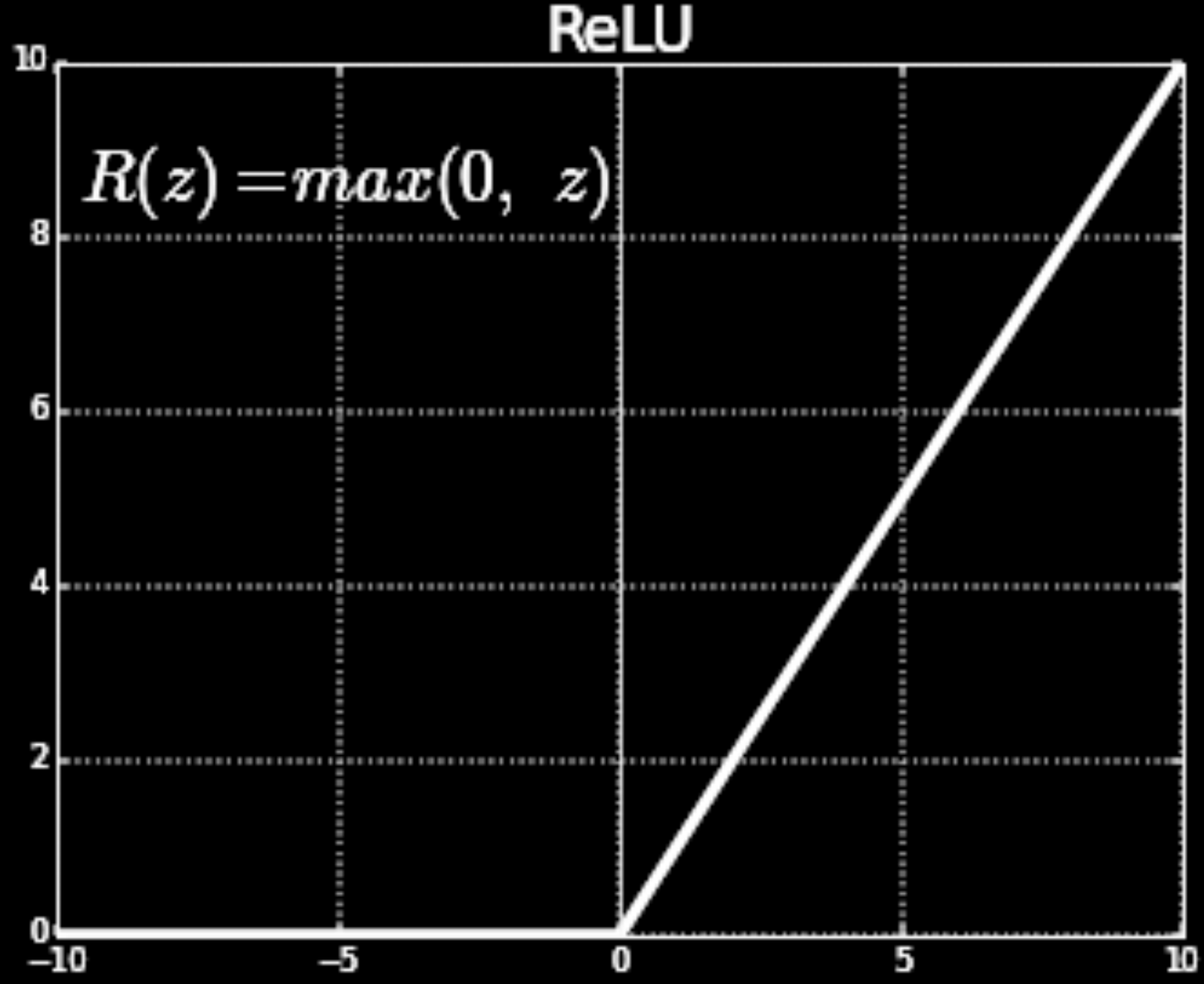
Background

- A neural network is a function with trainable parameters that learns a given mapping
  - Given an image, classify it as a cat or dog
  - Given a review, classify it as good or bad
  - Given a file, classify it as malware or benign

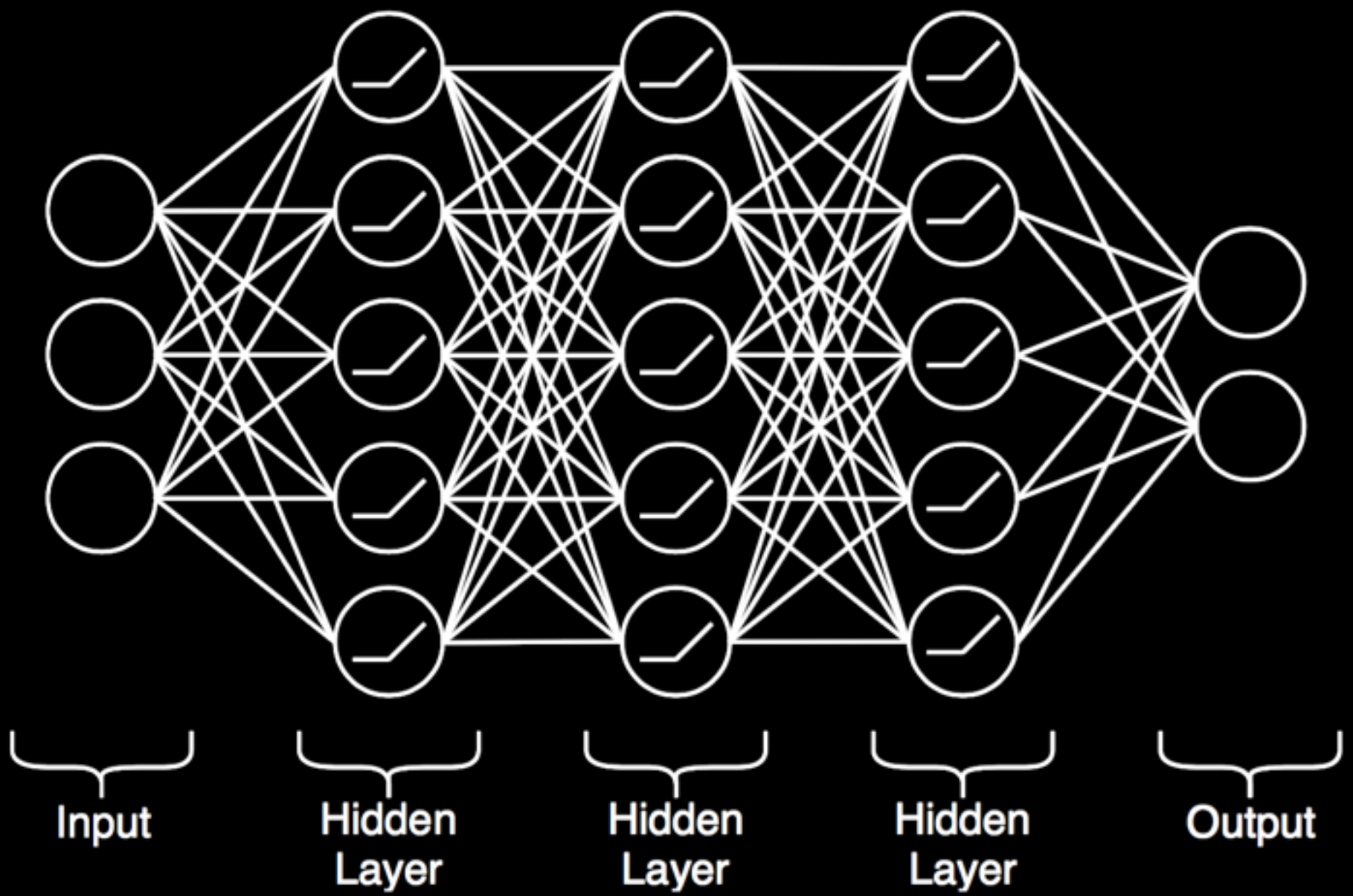
Background


d
i
n
a
e
j
o
s
b
f
k
p
t
c
g
l
q
h
m
r
Input
Hidden Layer
Hidden Layer
Hidden Layer
Output

Background



Background



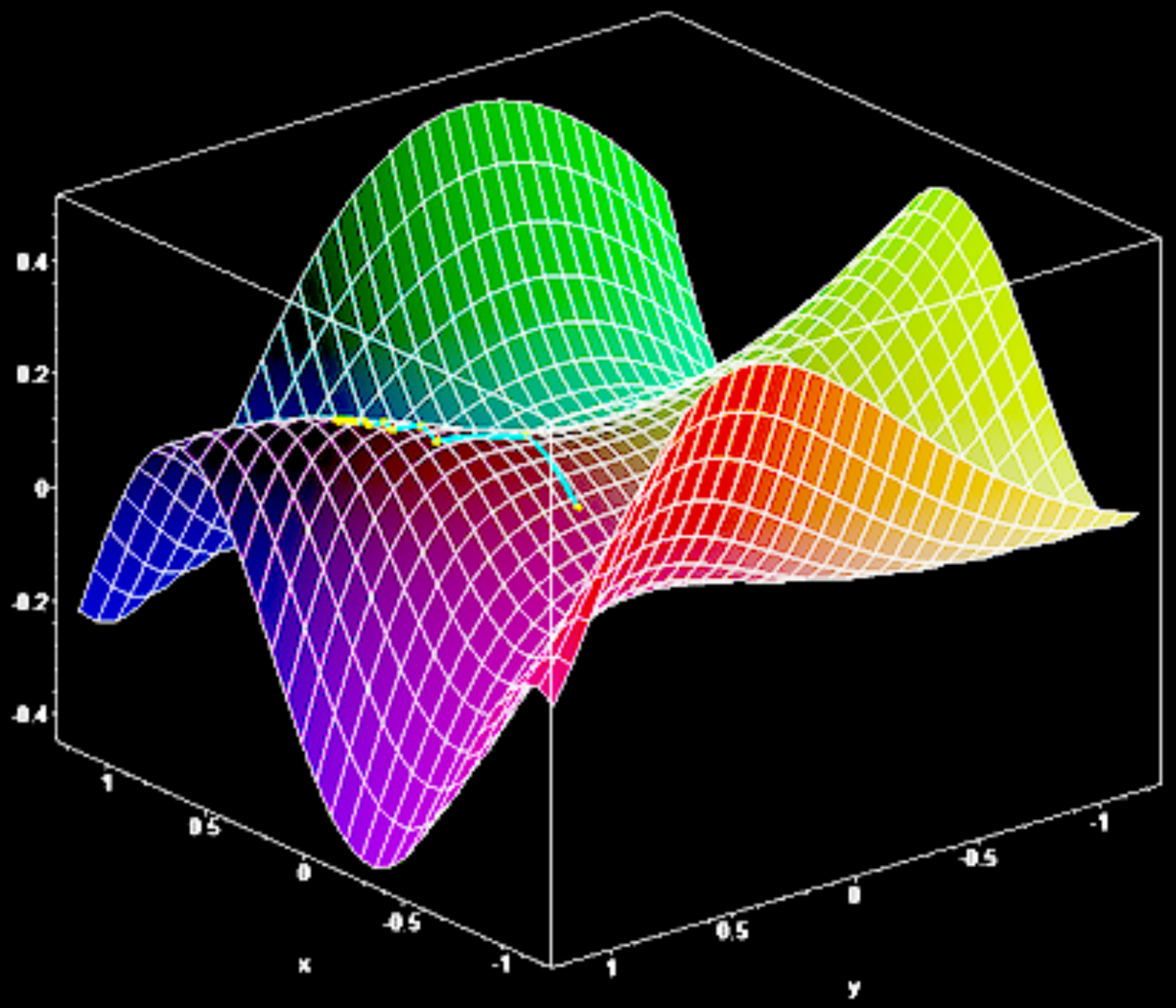
Background

- The output of a neural network F(x) is a probability distribution (p,q,...) where
  - p is the probability of class 1
  - q is the probability of class 2
  - ...

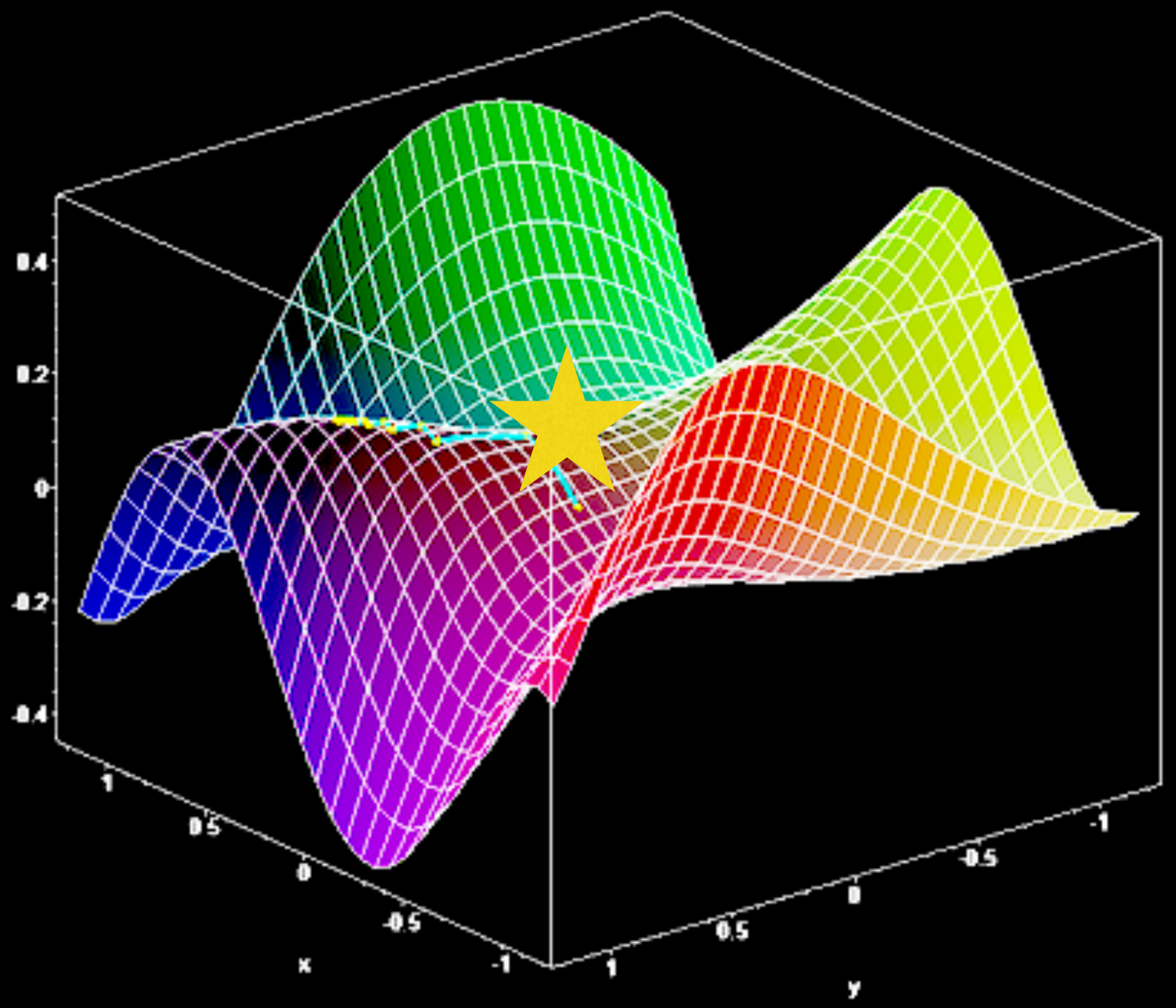
"Loss Function"

Measure of how accurate the network is

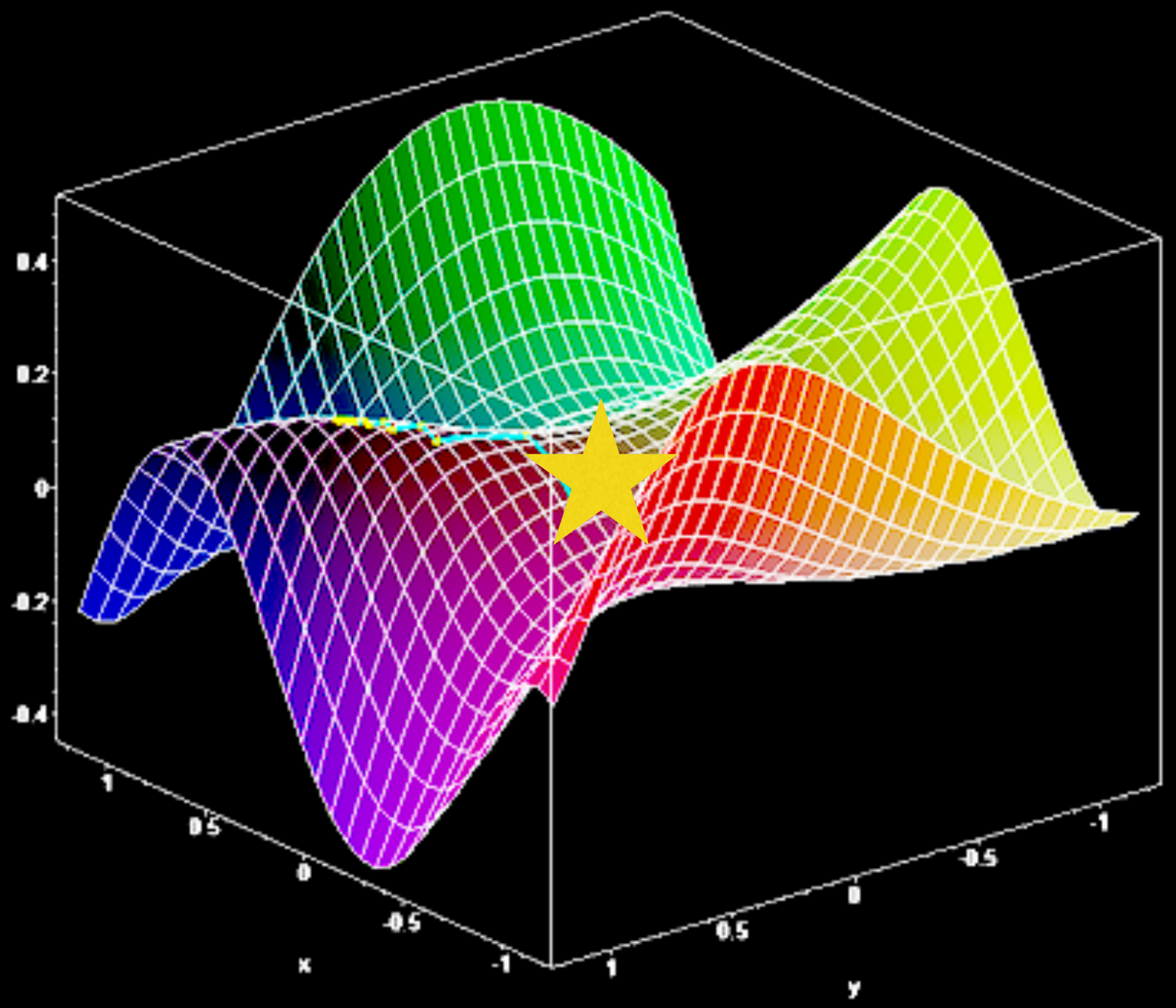
Background: gradient descent



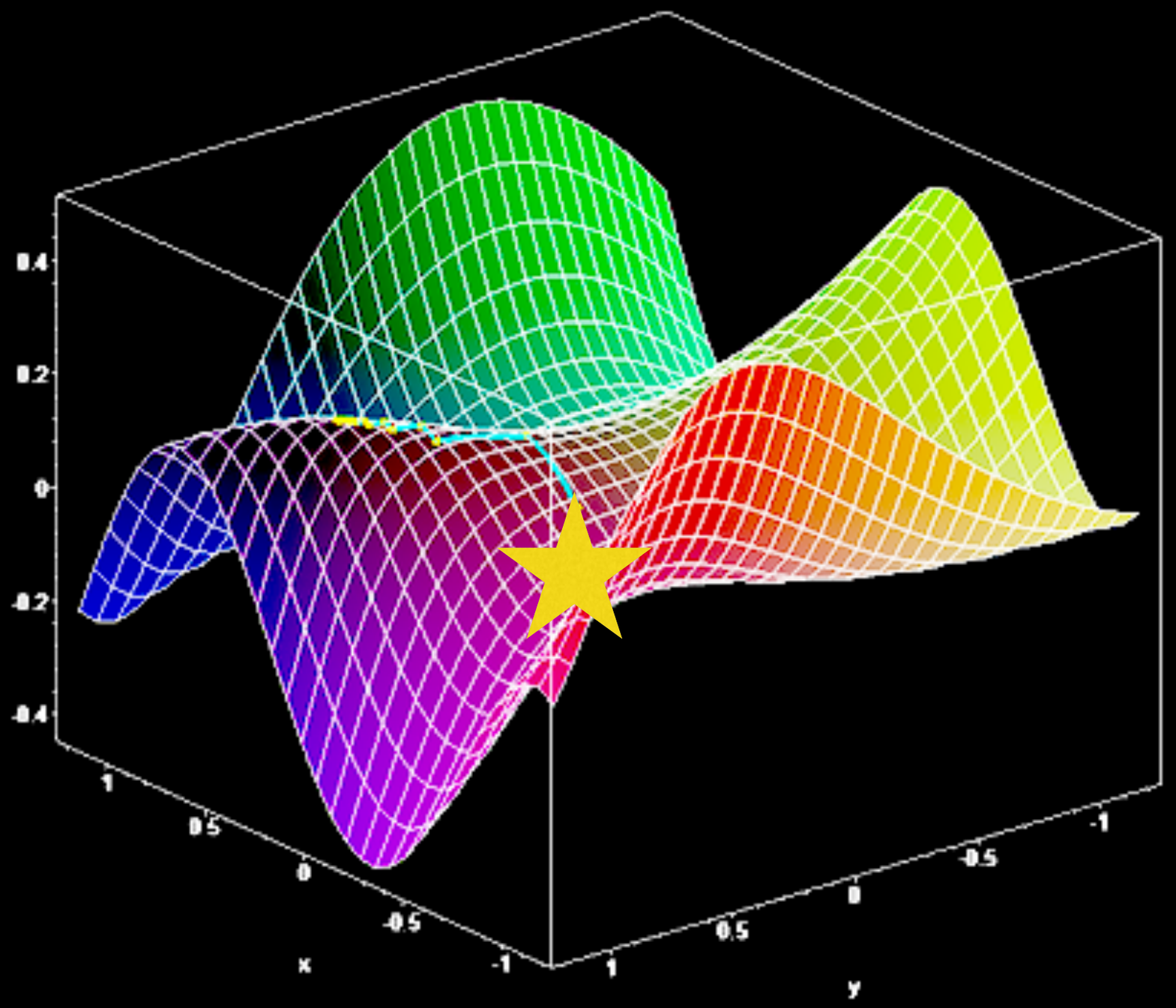
Background: gradient descent



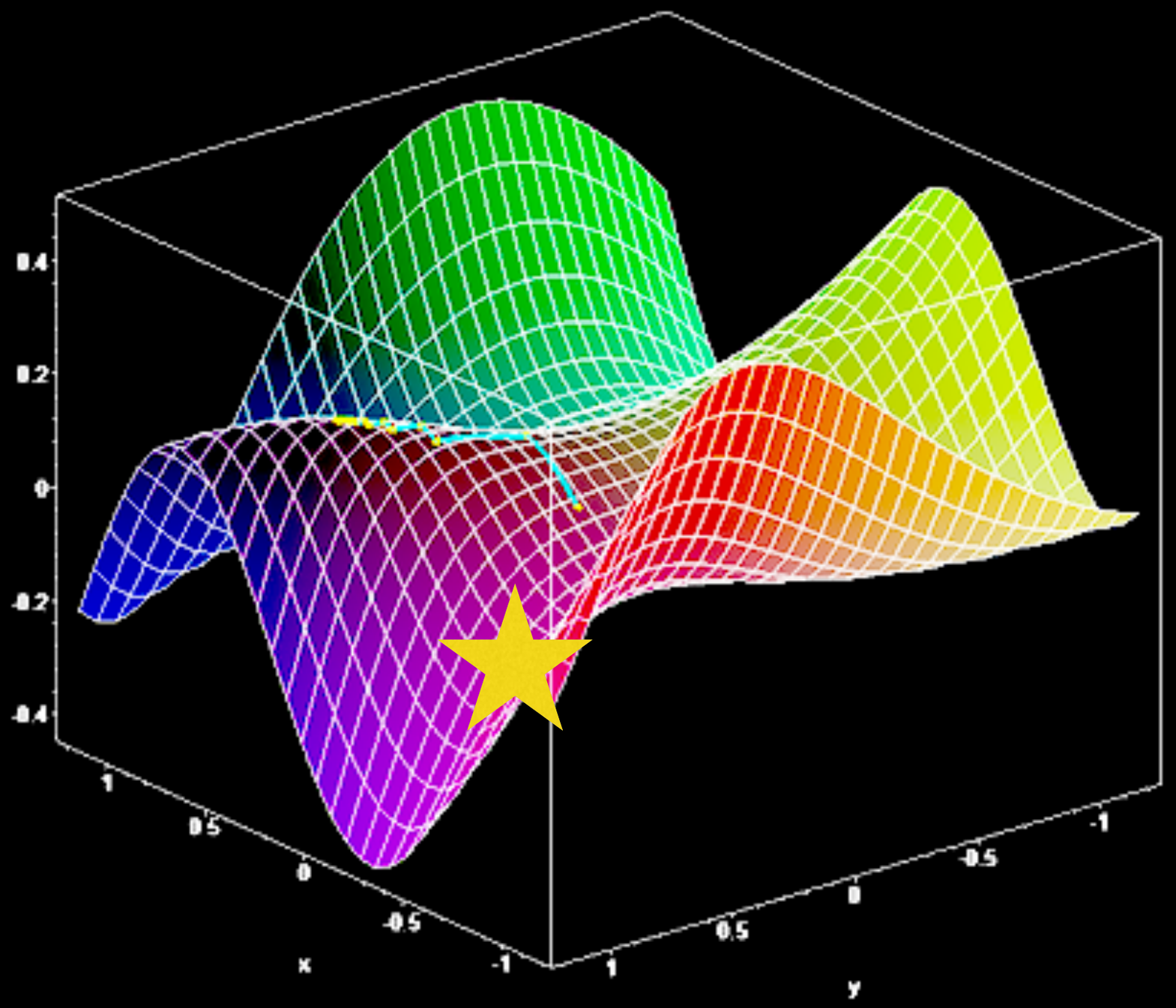
Background: gradient descent



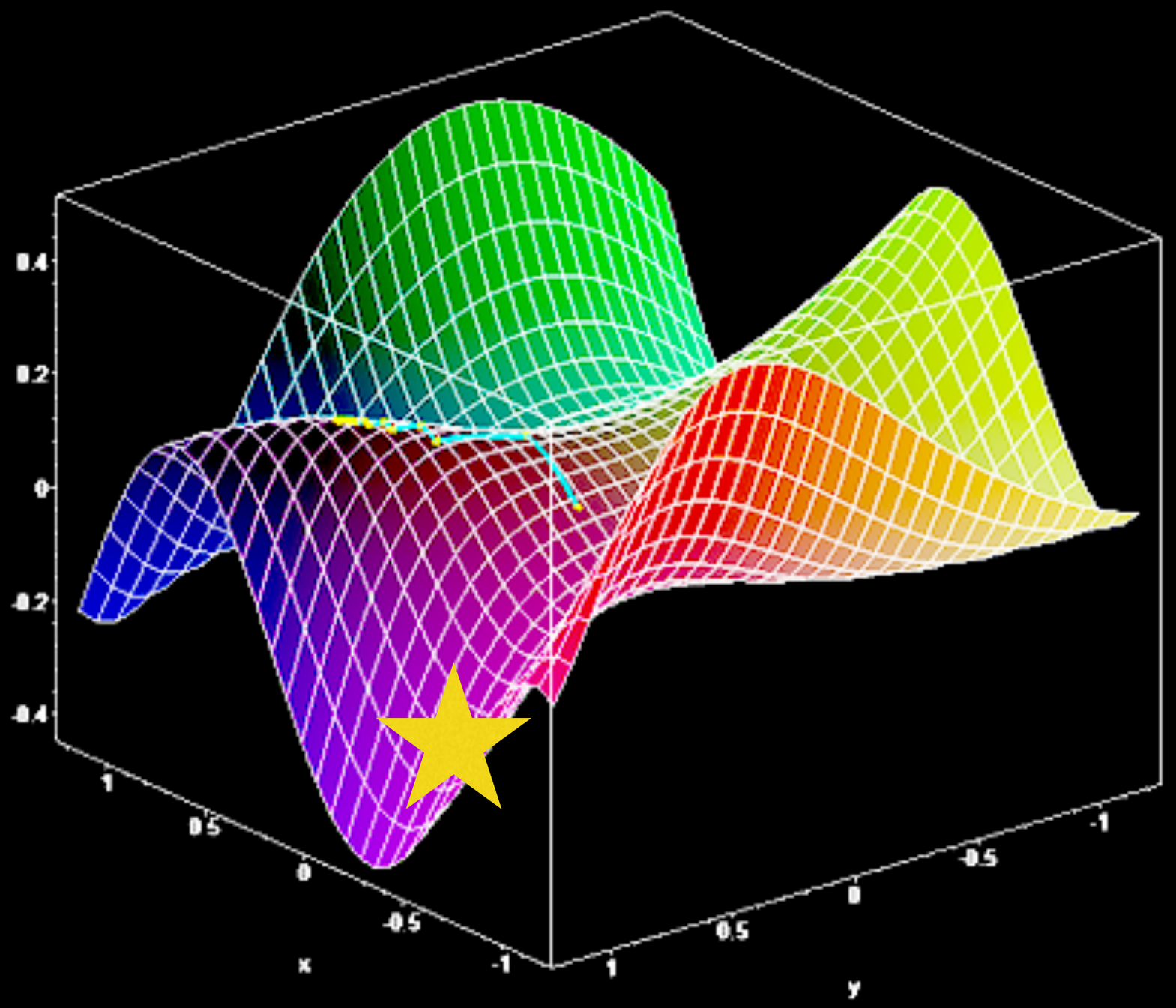
Background: gradient descent



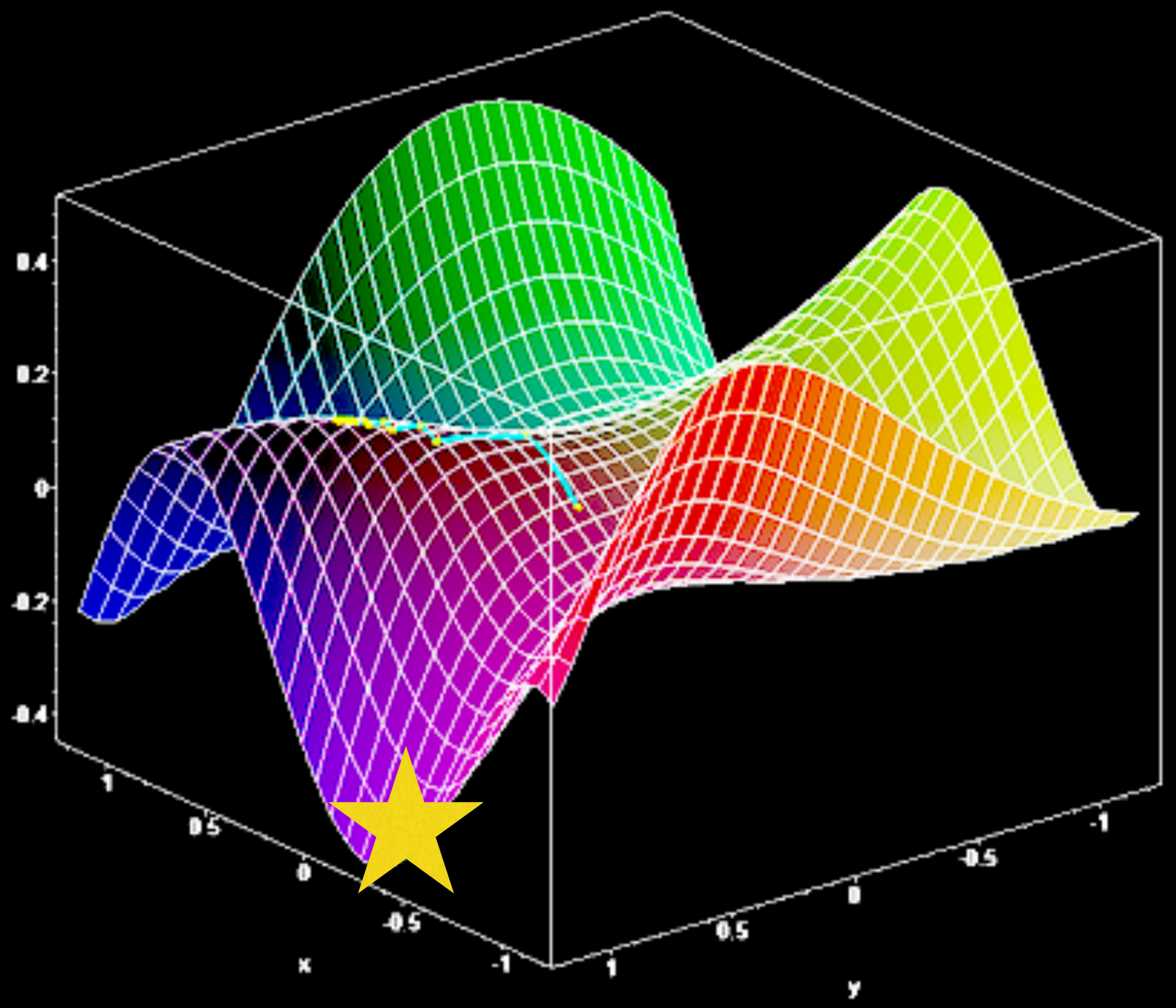
Background: gradient descent



Background: gradient descent



Background: gradient descent



Two important things:

1. Highly Non-Linear
2. Gradient Descent

ImageNet

Background: accuracy

- ImageNet 2011 best result: 75% accuracy
  No Neural Nets Used

- ImageNet 2012 best result: 85% accuracy
  Only top submission uses Neural Nets

- ImageNet 2013 best result: 89% accuracy
  ALL top submissions use Neural Nets

Best accuracy today:

97% accuracy

... but there's a catch

Background: Adversarial Examples

- Given an input X, and **any** label T ...
- ... it is easy to find an X′ close to X
- ... so that F(X′) = T

Dog

Hummingbird

Threat Model

- Adversary has access to model parameters
- Goal: construct adversarial examples

Defending Against Adversarial Examples

Huang, R., Xu, B., Schuurmans, D., and Szepesva´ri, C. Learning with a strong adversary. CoRR, abs/1511.03034 (2015)

Jin, J., Dundar, A., and Culurciello, E. Robust convolutional neural networks under adversarial noise. arXiv preprint arXiv:1511.06306 (2015)

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy (2016)

Hendrycks, D., and Gimpel, K. Visible progress on adversarial images and a new saliency map. arXiv preprint arXiv:1608.00530 (2016)

Li, X., and Li, F. Adversarial examples detection in deep networks with convolutional filter statistics. arXiv preprint arXiv:1612.07767 (2016)

Wang, Q. et al. Using Non-invertible Data Transformations to Build Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1610.01934 (2016).

Ororbia, I. I., et al. Unifying adversarial training algorithms with flexible deep data gradient regularization. arXiv preprint arXiv:1601.07213 (2016).

Wang, Q. et al. Learning Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1612.01401 (2016).

Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)

Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017)

Feinman, R., Curtin, R. R., Shintre, S., Gardner, A. B. Detecting Adversarial Samples from Artifacts. arXiv preprint arXiv:1703.00410 (2017)

Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and Clean Data Are Not Twins. arXiv preprint arXiv:1704.04960 (2017)

Dan Hendrycks and Kevin Gimpel. Early Methods for Detecting Adversarial Images. In International Conference on Learning Representations (Workshop Track) (2017)

Bhagoji, A. N., Cullina, D., and Mittal, P. Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. arXiv preprint arXiv:1704:02654 (2017)

Abbasi, M., and Christian G.. Robustness to Adversarial Examples through an Ensemble of Specialists. arXiv preprint arXiv:1702.06856 (2017).

Lu, J., Theerasit I., and David F. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. arXiv preprint arXiv:1704.00103 (2017)

Xu, W., Evans, D., and Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. arXiv preprint arXiv:1704.01155 (2017)

Hendrycks, D, and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv preprint arXiv:1610.02136 (2016)

Gondara, Lovedeep. Detecting Adversarial Samples Using Density Ratio Estimates. arXiv preprint arXiv:1705.02224 (2017)

Hosseini, Hossein, et al. Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318 (2017)

Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi. DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples. In ICLR (Workshop Track) (2017)

Wang, Q. et al. Adversary Resistant Deep Neural Networks with an Application to Malware Detection. arXiv preprint arXiv:1610.01239 (2017)

Cisse, Moustapha, et al. Parseval Networks: Improving Robustness to Adversarial Examples. arXiv preprint arXiv:1704.08847 (2017).

Nayebi, Aran, and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. arXiv preprint arXiv:1703.09202 (2017).

Defending Against Adversarial Examples

Huang, R., Xu, B., Schuurmans, D., and Szepesva´ri, C. Learning with a strong adversary. CoRR, abs/1511.03034 (2015)

Jin, J., Dundar, A., and Culurciello, E. Robust convolutional neural networks under adversarial noise. arXiv preprint arXiv:1511.06306 (2015)

Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. IEEE Symposium on Security and Privacy (2016)

Hendrycks, D., and Gimpel, K. Visible progress on adversarial images and a new saliency map. arXiv preprint arXiv:1608.00530 (2016)

Li, X., and Li, F. Adversarial examples detection in deep networks with convolutional filter statistics. arXiv preprint arXiv:1612.07767 (2016)

Wang, Q. et al. Using Non-invertible Data Transformations to Build Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1610.01934 (2016).

Ororbia, I. I., et al. Unifying adversarial training algorithms with flexible deep data gradient regularization. arXiv preprint arXiv:1601.07213 (2016).

Wang, Q. et al. Learning Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1612.01401 (2016).

Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)

Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017)

Feinman, R., Curtin, R. R., Shintre, S., Gardner, A. B. Detecting Adversarial Samples from Artifacts. arXiv preprint arXiv:1703.00410 (2017)

Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and Clean Data Are Not Twins. arXiv preprint arXiv:1704.04960 (2017)

Dan Hendrycks and Kevin Gimpel. Early Methods for Detecting Adversarial Images. In International Conference on Learning Representations (Workshop Track) (2017)

Bhagoji, A. N., Cullina, D., and Mittal, P. Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. arXiv preprint arXiv:1704:02654 (2017)

Abbasi, M., and Christian G.. Robustness to Adversarial Examples through an Ensemble of Specialists. arXiv preprint arXiv:1702.06856 (2017).

Lu, J., Theerasit I., and David F. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. arXiv preprint arXiv:1704.00103 (2017)

Xu, W., Evans, D., and Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. arXiv preprint arXiv:1704.01155 (2017)

Hendrycks, D, and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv preprint arXiv:1610.02136 (2016)

Gondara, Lovedeep. Detecting Adversarial Samples Using Density Ratio Estimates. arXiv preprint arXiv:1705.02224 (2017)

Hosseini, Hossein, et al. Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318 (2017)

Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi. DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples. In ICLR (Workshop Track) (2017)

Wang, Q. et al. Adversary Resistant Deep Neural Networks with an Application to Malware Detection. arXiv preprint arXiv:1610.01239 (2017)

Cisse, Moustapha, et al. Parseval Networks: Improving Robustness to Adversarial Examples. arXiv preprint arXiv:1704.08847 (2017).

Nayebi, Aran, and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. arXiv preprint arXiv:1703.09202 (2017).

This talk:

How should we evaluate if a defense to adversarial examples is effective?

Two ways to evaluate robustness:

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

Key Insight #1:

Gradient descent works very well for training neural networks. Why not for breaking them too?

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′) <= 0 if F(x′) = T

  - g(x′) > 0 if F(x′) != T

Reformulation

- For example
  - $g(x') = (1-F(x')_T)$
- If $F(x')$ says the probability of T is 1:
  - $g(x') = (1-F(x')_T) = (1-1) = 0$
- $F(x')$ says the probability of T is 0:
  - $g(x') = (1-F(x')_T) = (1-0) = 1$

Key Insight #2:

The loss function you choose is important

... so, is this approach good?

Evaluation

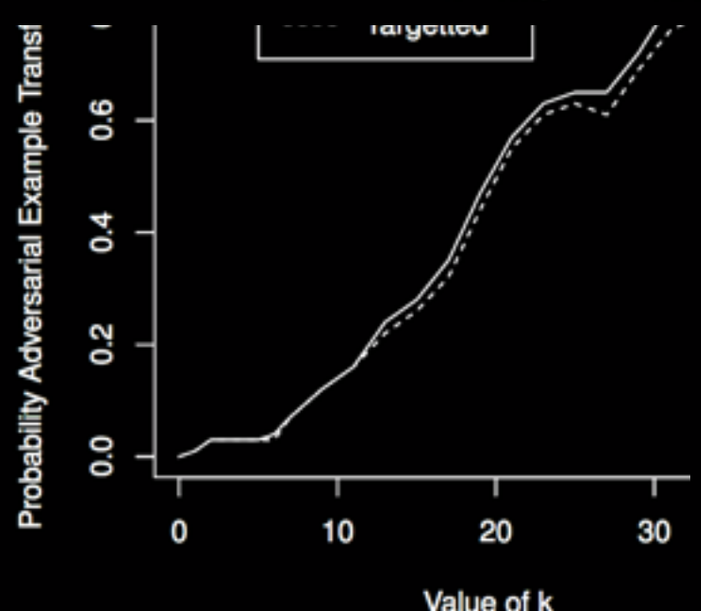
Parameter	MNIST Model	CIFAR Model
Learning Rate	0.1	0.01 (decay 0.5)
Momentum	0.9	0.9 (decay 0.5)
Delay Rate	-	10 epochs
Dropout	0.5	0.5
Batch Size	128	128
Epochs	50	50

TABLE II

	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.



	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.

del

Fully Connected + ReLU	200	256
Softmax	10	10

Mo
AI

THIS
VE

	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%
			0.96	100%	2.22	100%
			-	-	-	-
			006	100%	0.01	100%
			064	2%	-	0%
			0.01	99%	0.03	98%

10
tance

Evaluation #1:
Comparing to Other Attacks

Original

Previous Attack

Our Attack

Dog

Hummingbird

Dog

Hummingbird

Dog
(83%)

Hummingbird
(98%)

Evaluation #2:
Breaking Current Defenses

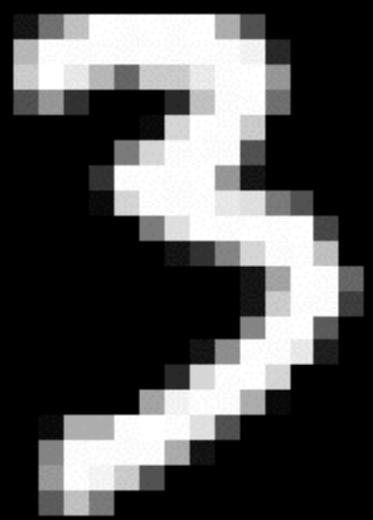
Our attacks defeat the strongest defense.

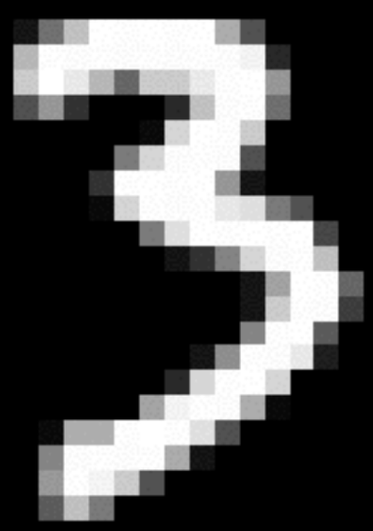
Distillation as a defense to adversarial perturbations against deep neural networks.
Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. IEEE S&P (2016)

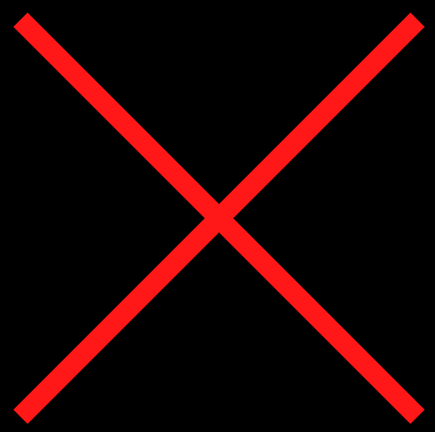
Original

Previous Attack

Our Attack







So I'm Building A Defense. What Should I Do To Evaluate It?

- Release your source code
  - This is an empirical science
- Evaluate against the strongest attack as a baseline
  - Robustness against weak attacks is useless

https://nicholas.carlini.com/code/nn_robust_attacks/

Backup Slides

Dog

Hummingbird

Broken Defenses

Huang, R., Xu, B., Schuurmans, D., and Szepesva´ri, C. Learning with a strong adversary. CoRR, abs/1511.03034 (2015)

Jin, J., Dundar, A., and Culurciello, E. Robust convolutional neural networks under adversarial noise. arXiv preprint arXiv:1511.06306 (2015)

~~Papernot, N., McDaniel, P., Wu, X., Jha, S., and Swami, A. Distillation as a defense to adversarial perturbations against deep neural networks. IEEE S&P (2016)~~

Hendrycks, D., and Gimpel, K. Visible progress on adversarial images and a new saliency map. arXiv preprint arXiv:1608.00530 (2016)

~~Li, X., and Li, F. Adversarial examples detection in deep networks with convolutional filter statistics. arXiv preprint arXiv:1612.07767 (2016)~~

Wang, Q. et al. Using Non-invertible Data Transformations to Build Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1610.01934 (2016).

Ororbia, I. I., et al. Unifying adversarial training algorithms with flexible deep data gradient regularization. arXiv preprint arXiv:1601.07213 (2016).

Wang, Q. et al. Learning Adversary-Resistant Deep Neural Networks. arXiv preprint arXiv:1612.01401 (2016).

~~Grosse, K., Manoharan, P., Papernot, N., Backes, M., and McDaniel, P. On the (statistical) detection of adversarial examples. arXiv preprint arXiv:1702.06280 (2017)~~

~~Metzen, J. H., Genewein, T., Fischer, V., and Bischoff, B. On detecting adversarial perturbations. arXiv preprint arXiv:1702.04267 (2017)~~

~~Feynman, R., Curtin, R. R., Shintre, S., Gardner, A. B. Detecting Adversarial Samples from Artifacts. arXiv preprint arXiv:1703.00410 (2017)~~

~~Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku. Adversarial and Clean Data Are Not Twins. arXiv preprint arXiv:1704.04960 (2017)~~

~~Dan Hendrycks and Kevin Gimpel. Early Methods for Detecting Adversarial Images. In International Conference on Learning Representations (Workshop Track) (2017)~~

~~Bhagoji, A. N., Cullina, D., and Mittal, P. Dimensionality Reduction as a Defense against Evasion Attacks on Machine Learning Classifiers. arXiv preprint arXiv:1704:02654 (2017)~~

Abbasi, M., and Christian G.. Robustness to Adversarial Examples through an Ensemble of Specialists. arXiv preprint arXiv:1702.06856 (2017).

Lu, J., Theerasit I., and David F. SafetyNet: Detecting and Rejecting Adversarial Examples Robustly. arXiv preprint arXiv:1704.00103 (2017)

Xu, W., Evans, D., and Qi, Y. Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks. arXiv preprint arXiv:1704.01155 (2017)

Hendrycks, D, and Gimpel, K. A Baseline for Detecting Misclassified and Out-of-Distribution Examples in Neural Networks. arXiv preprint arXiv:1610.02136 (2016)

Gondara, Lovedeep. Detecting Adversarial Samples Using Density Ratio Estimates. arXiv preprint arXiv:1705.02224 (2017)

Hosseini, Hossein, et al. Blocking transferability of adversarial examples in black-box learning systems. arXiv preprint arXiv:1703.04318 (2017)

Ji Gao, Beilun Wang, Zeming Lin, Weilin Xu, Yanjun Qi. DeepCloak: Masking Deep Neural Network Models for Robustness Against Adversarial Samples. In ICLR (Workshop Track) (2017)

Wang, Q. et al. Adversary Resistant Deep Neural Networks with an Application to Malware Detection. arXiv preprint arXiv:1610.01239 (2017)

Cisse, Moustapha, et al. Parseval Networks: Improving Robustness to Adversarial Examples. arXiv preprint arXiv:1704.08847 (2017).

Nayebi, Aran, and Surya Ganguli. Biologically inspired protection of deep networks from adversarial attacks. arXiv preprint arXiv:1703.09202 (2017).

	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.