IPAM Autonomous Vehicles

Regularization and Adversarial Robustness

Feb 25-March 1, 2019

Adam Oberman
McGill Dept of Math and Stats

supported by NSERC, AFOSR FA9550-18-1-0167

Image Classification by CNNs

Approximation Theory:

+

Random sampling:

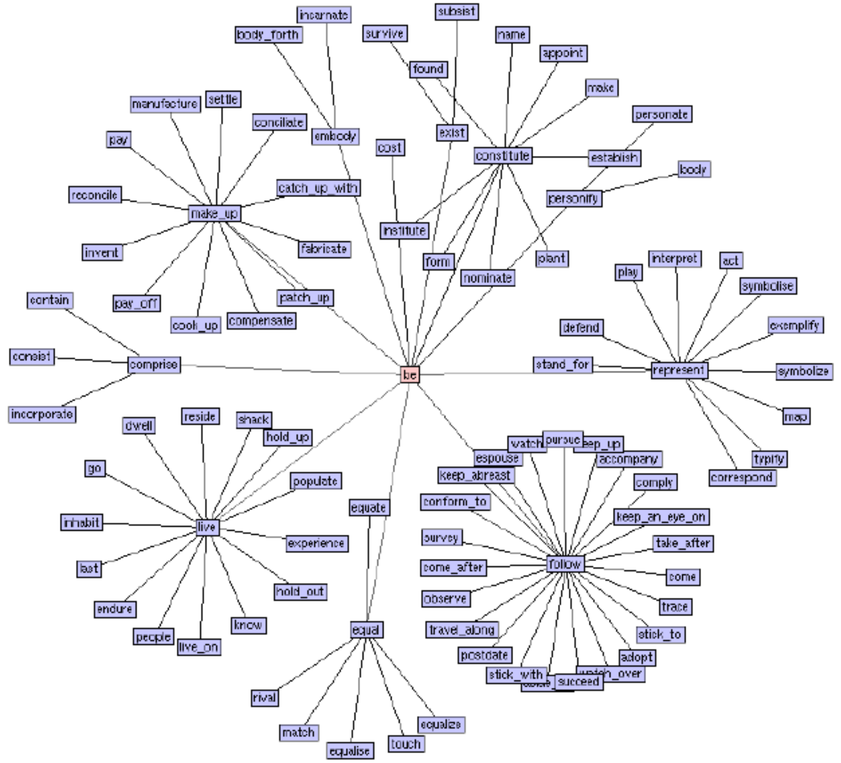
Fitting a map from images to labels

+

Images are sampled from distribution

x in M = manifold of images

f(x)



graph or list of word labels

ImageNet

- ImageNet: Total number of classes: m =21841
- Total number of images: n =14,197,122
- Color images d= 3*256*256= 196,608

Facebook used 256 GPUs, working in parallel, to train ImageNet.

Still an academic dataset. Total number of images on Facebook is much larger

We still don't understand why it works so well

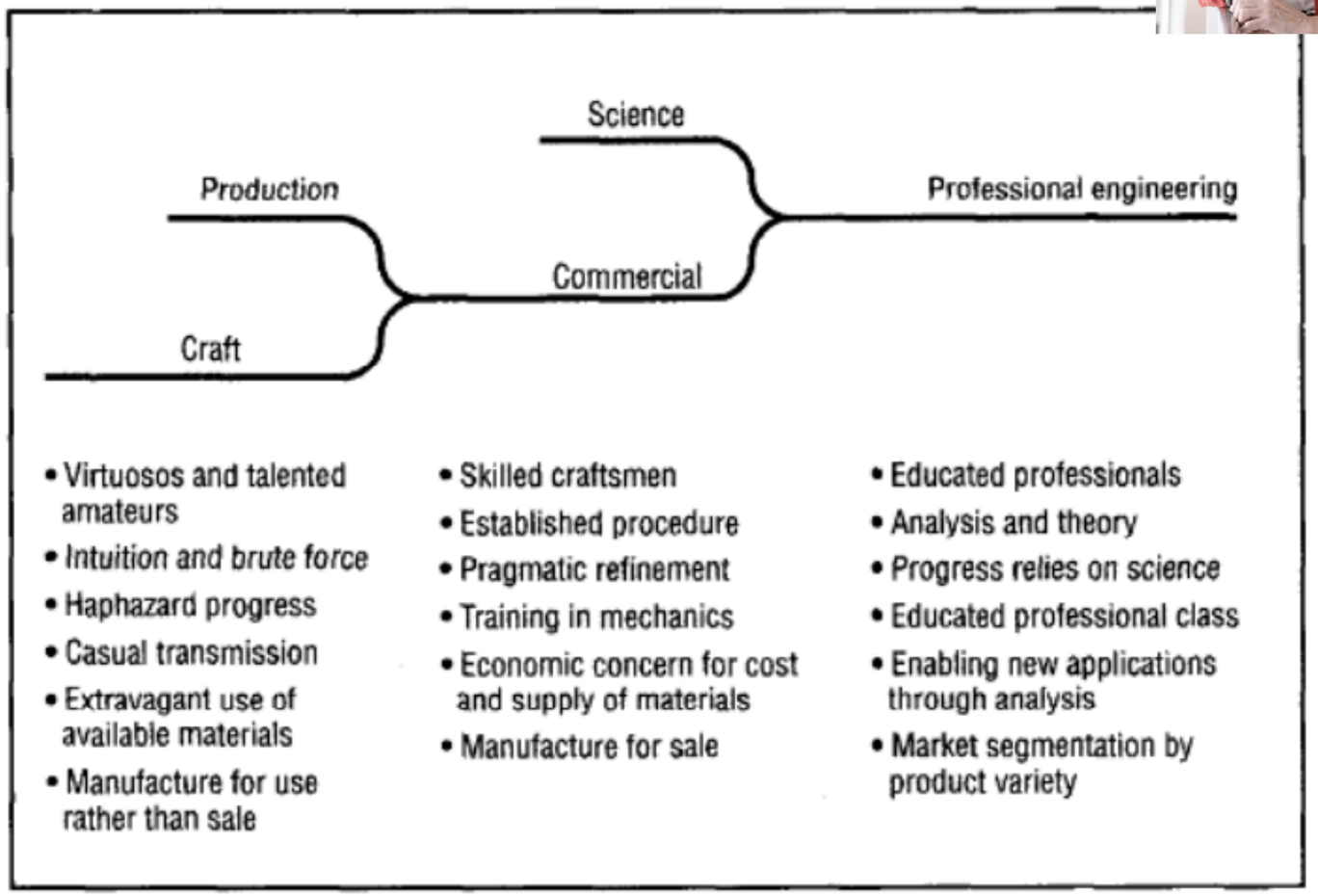
In theory, due to curse of dimensionality, ***impossible*** to accurately interpolate a high dimensional function.

In practice, ***possible*** using Deep Neural Network architecture, training to fit the data with SGD. However we don't know why it works.

Can train a computer to caption images more accurately that human performance.

Figure 1. At the current state-of-art, more than 95% of images can be c

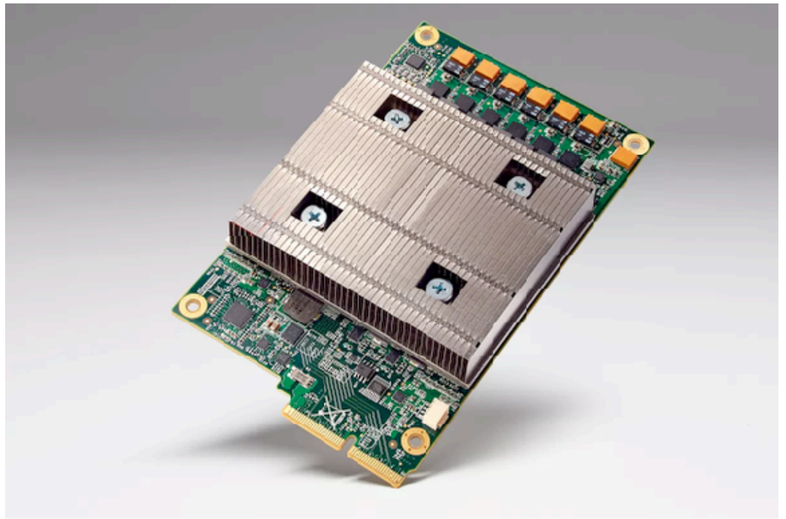
Mary Shaw's evolution of software engineering discipline



Better theory: improves reliability and discipline evolves

Are all models with the same generalization loss equally good?

Besides generalization/accuracy what else do we care about?



- Cost of model:
  - memory storage required,
  - inference time
  - power usage (in hardware)
- Robustness: sensitivity of the model to small changes in the data.
- Stability: sensitivity of the model to small changes in the model parameters

Challenges for deep learning

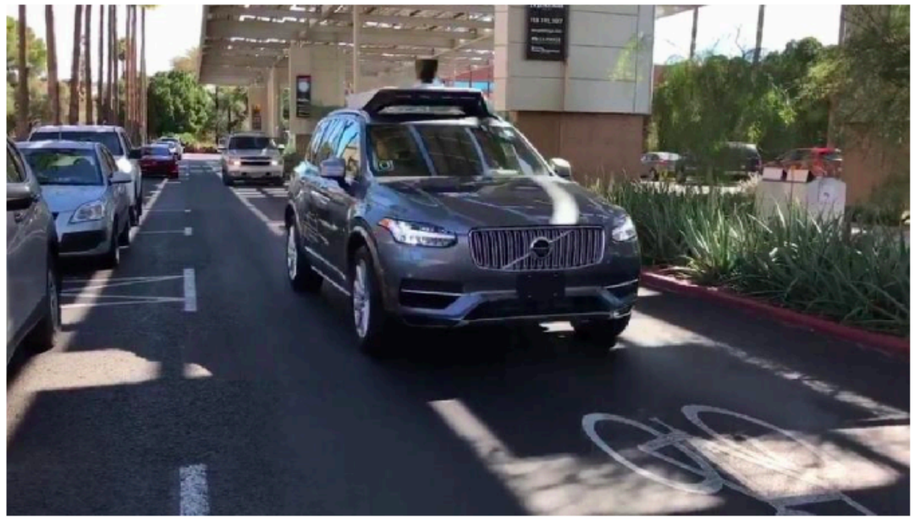
"It is not clear that the existing AI paradigm is immediately amenable to any sort of software engineering validation and verification. This is a serious issue, and is a potential roadblock to DoD's use of these modern AI systems, especially when considering the liability and accountability of using AI"

JASON report

Self-Driving Uber Hits, Kills Pedestrian in Arizona

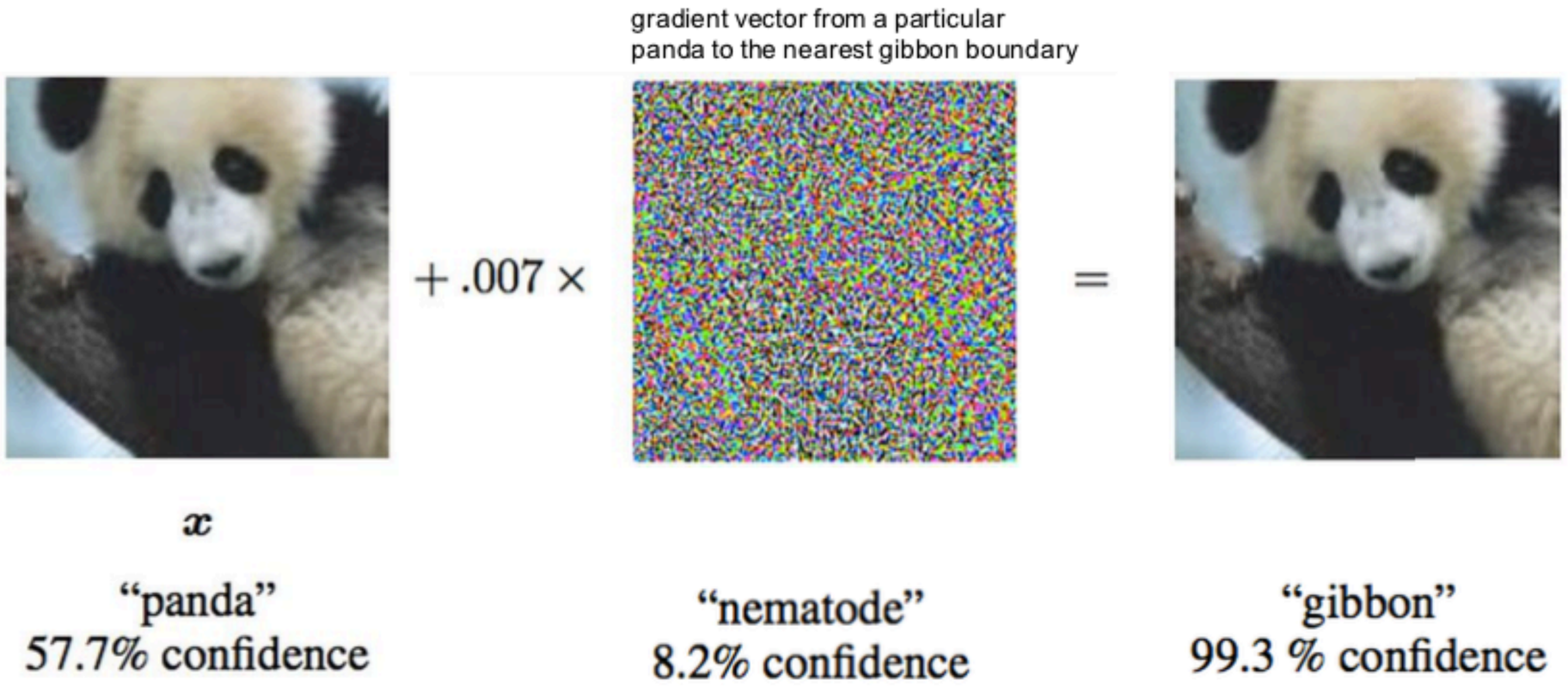
The Uber vehicle was operating in autonomous mode with a human behind the wheel in Tempe, Arizona, when the incident occurred overnight.

By Angela Moscaritolo March 19, 2018 2:07PM EST



Learning networks. Two things to make clear to the reader (1) We don't know how Deep Learning works and (2) when it makes a prediction, we don't have an explanation why it arrived at that prediction. That is just scratching the

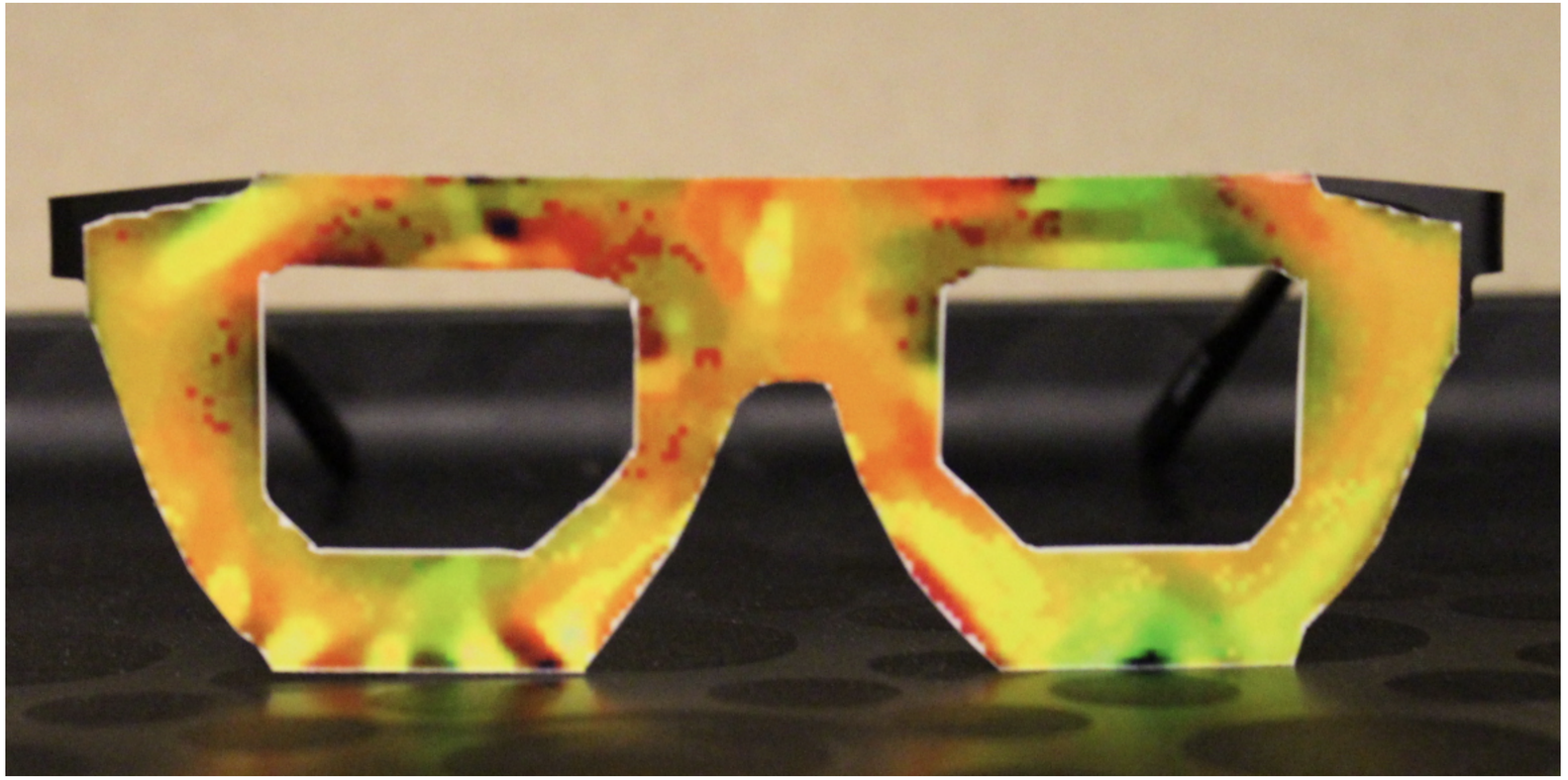
Intro: Adversarial Attacks



Small (visually imperceptible) perturbations of an image lead to misclassification

Source: EXPLAINING AND HARNESSING ADVERSARIAL EXAMPLES, Goodfellow

Attacks against Facial Recognition



Glasses make you invisible to facial recognition

Source: Accessorize to a Crime: Real and Stealthy Attacks on State-of-the-Art Face Recognition. Sharif et al.

Attacks on road sign classification

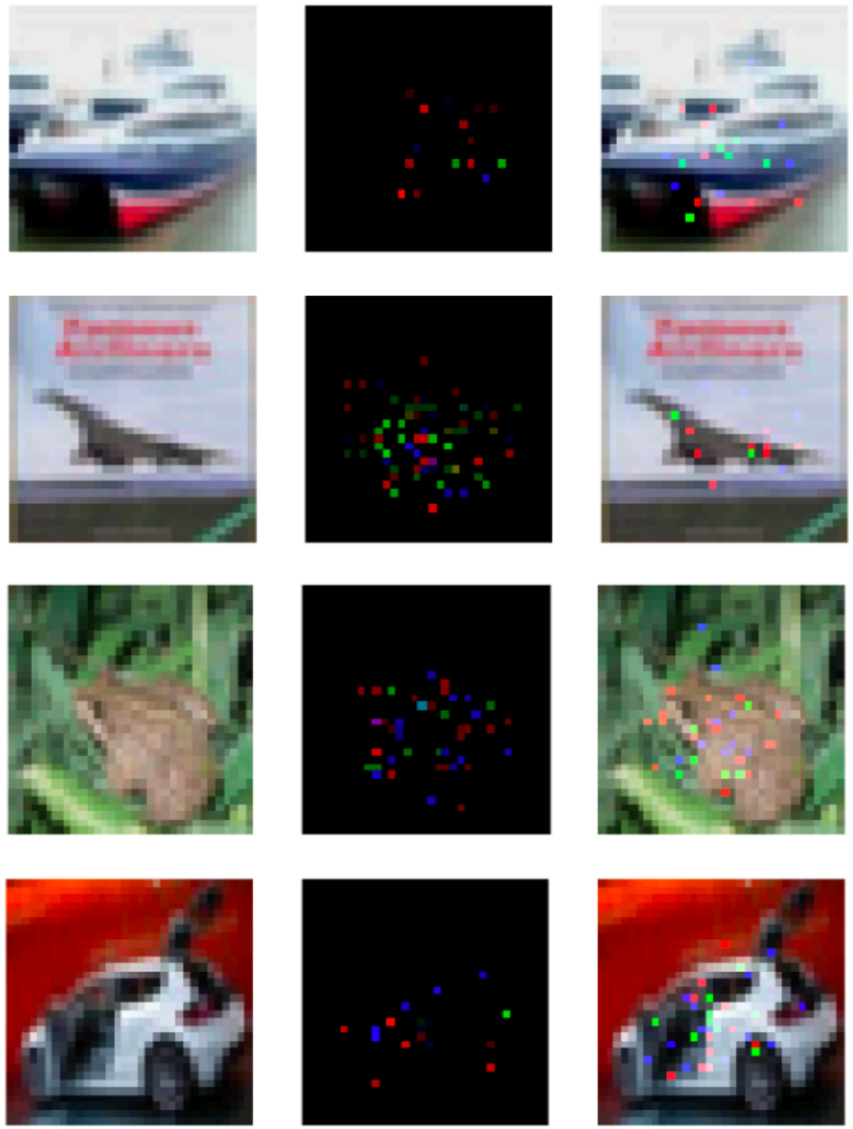
Left: real graffiti on a stop sign, something that most humans would not think is suspicious. Right: a physical perturbation applied to a stop sign. Models classify the sign on the right as a **Speed Limit: 45 mph sign**!

Source: Robust Physical-World Attacks on Deep Learning Visual Classification.

Jacobian Saliency Attack

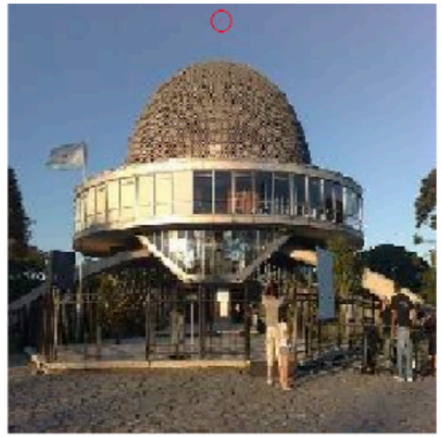
The Limitations of Deep Learning in Adversarial Settings Nicolas Papernot, (Vector Institute)

Use the Jacobian of the model to see which pixels are influential on classification, and change those.



One pixel attack

Can even choose just one of the pixels, and fool the network



Planetarium

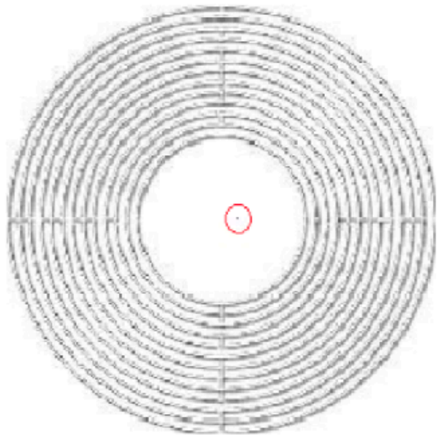
Mosque(7.81%)

Comforter

Pillow(6.83%)

Jellyfish

Bathing tub(21.18%)



Whorl

Blower (37.00%)

One pixel attack for fooling deep neural networks
Jiawei Su,

Scale measures visible attacks

DNNs are vulnerable to attacks which are invisible to the human eye.
Undefended networks have 100% error rate at .1 (in max norm)

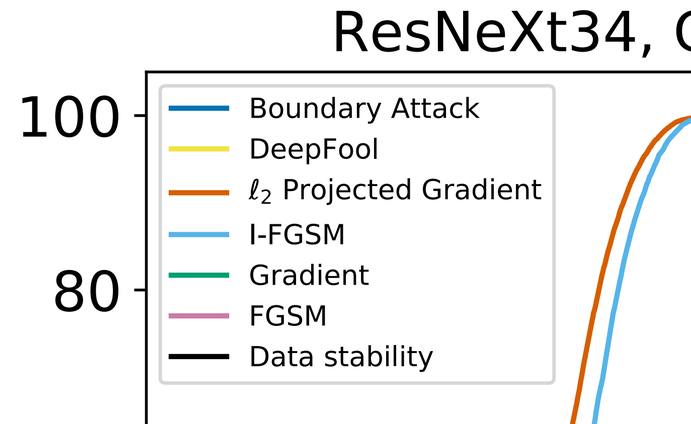
Adversarial attacks, defence and detection

- Why are models robust to random perturbations, but not adversarial ones?
- Designing adversarial attacks on a given network is an **optimization** problem: find the smallest perturbation of an image which leads to misclassification.
- Adversarial defences (models which resist attacks)
  - **Game Theory** problem: need to anticipate the attacks, and defend against them. It matters who moves first
  - Also a **Security** problem: like spam detection and encryption, need to be aware of possible attacks and defend against them. Practical considerations (time, effort) matter as much as theoretical ones (security guarantees).
  - Address it using **Regularization** coming from calculus of variations

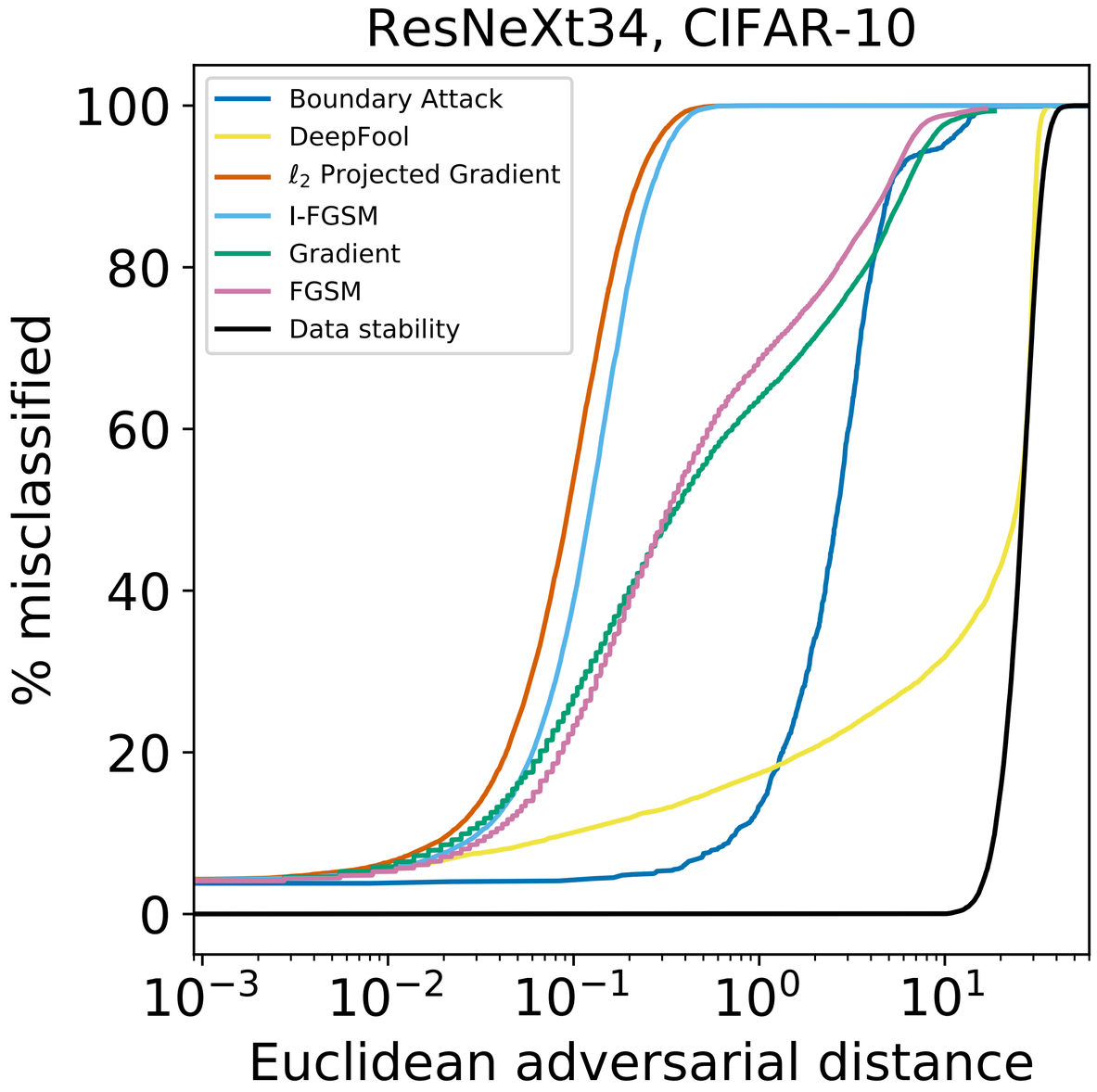
Foolbox Attacks

- Benchmark attacks.
- White box / Black box (gradients of model, just model predictions)
- Gradient based attacks on the loss

Foolbox: A Python toolbox to benchmark the robustness of machine learning models
Jonas Rauber, Wieland Brendel, Matthias Bethge



Arms race of attack methods and defences



Error curve: *probability an image misclassified as a function of adversarial attack vector norm.*

Error curve for an undefended model for different attacks.

Strongest attacks on undefended model:

Iterative Gradient Attacks (FGSM for infinity norm/Projected Gradient for Euclidean norm)

Attack Detection

- Can't defend very well, but can you detect attacks.
- 8 papers published in 2017 conferences, detecting attacks
- *Game theory*: detector moves after attacker.

Carlini-Wagner Evasion Attacks

- Then Carlini-Wagner came and broke every detection method.
- They used a modifies loss function, which used knowledge of the detection method to optimize
  - misclassification + undetectable
- The strongest model was a Bayesian combination of 30 models which used consensus. Harder to fool 30 models. (But costs more)
- *Game Theory*: Carlini moves after detector.

Nicolas Carlini

Gradient Obfuscation and the Arms Race

- Strongest attacks on undefended models are gradient based
- On many models, Carlini-Wagner and Boundary attacks are 1000 times slower, and not as effective, because they do not use gradient information.
- However, early defended models seemed to work well against gradient attacks. (But they didn't check the black box attacks, which were assumed to be weaker).
- Turns out: non-gradient based attacks destroyed them.
- Carlini-Wagner called these models "gradient obfuscation", essentially providing bad gradients.
- Whenever a gradient attack works better than a black box attack they call it obfuscated.
- **Lesson**: need to test all the attacks.

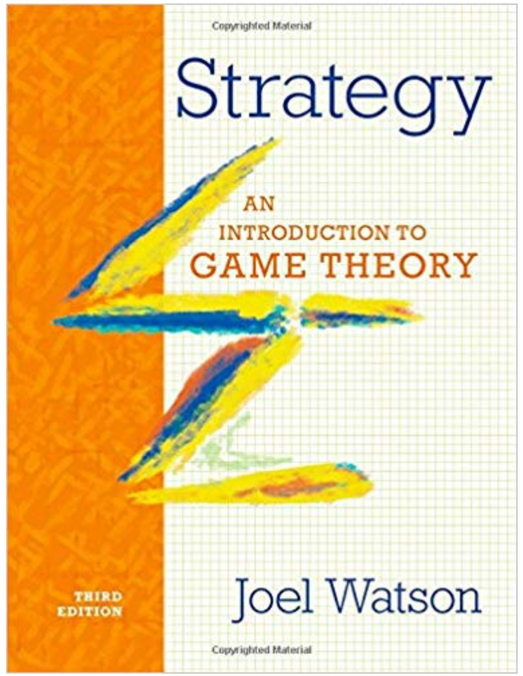
Madry: Defence by adversarial training

- Simple idea: train network replacing original images with attacked images (still using correct labels).
- Now when someone attacks the images, the model has already been trained to recognize them.
- Benefits: improved adversarial robustness
- Problem: loss of accuracy (say from 4% to 12% on CIFAR 10).
  - Madry: "Robustness may be at odds with accuracy" claims this loss is unavoidable
- Algorithmic / Variational :
  - Our interpretation: Adversarial Training corresponds to Total Variation Regularization (to be explained)
  - Better results: implement the Regularization effectively.

Aleksander Mądry

Lessons from Game Theory

- Declaring your strategy is a disadvantage / moving after you see the other player's move is an advantage.
- Custom defences may work well if the attack is known
- New attacks can take advantage of the knowledge
- Min Max strategy is best outcome over the worst attack : no disadvantage to declaring strategy
- **Game theory lesson:** Rock Paper Scissors: R > S, S > P, but R < P (!)
- **Game theory lesson:** Rock Paper Scissors: minimax strategy is randomized.



Security Lessons

- Encryption : there may be no unbreakable code. However can make a code require so much effort to break, that impractical (e.g. Prime Factorization)
- Bank vault: takes long time to break in.
- Spam Detection : will always get false positive / false negatives
  - sometimes real message classified as spam (false positive)
  - sometimes miss spam, classify as ham (false negative)
  - There is always a trade off between these.
  - Better tests may require more effort: e.g. medical screenings. Less invasive screen, then more invasive test.

Adversarial Robustness without loss of accuracy.

Chris Finlay (current PhD student)

Bilal Abbasi (former PhD now working in AI)

Improved robustness to adversarial examples using Lipschitz regularization of the loss

Chris Finlay, O., Bilal Abbasi; Oct 2018; arxiv

Outcome of the regularized model

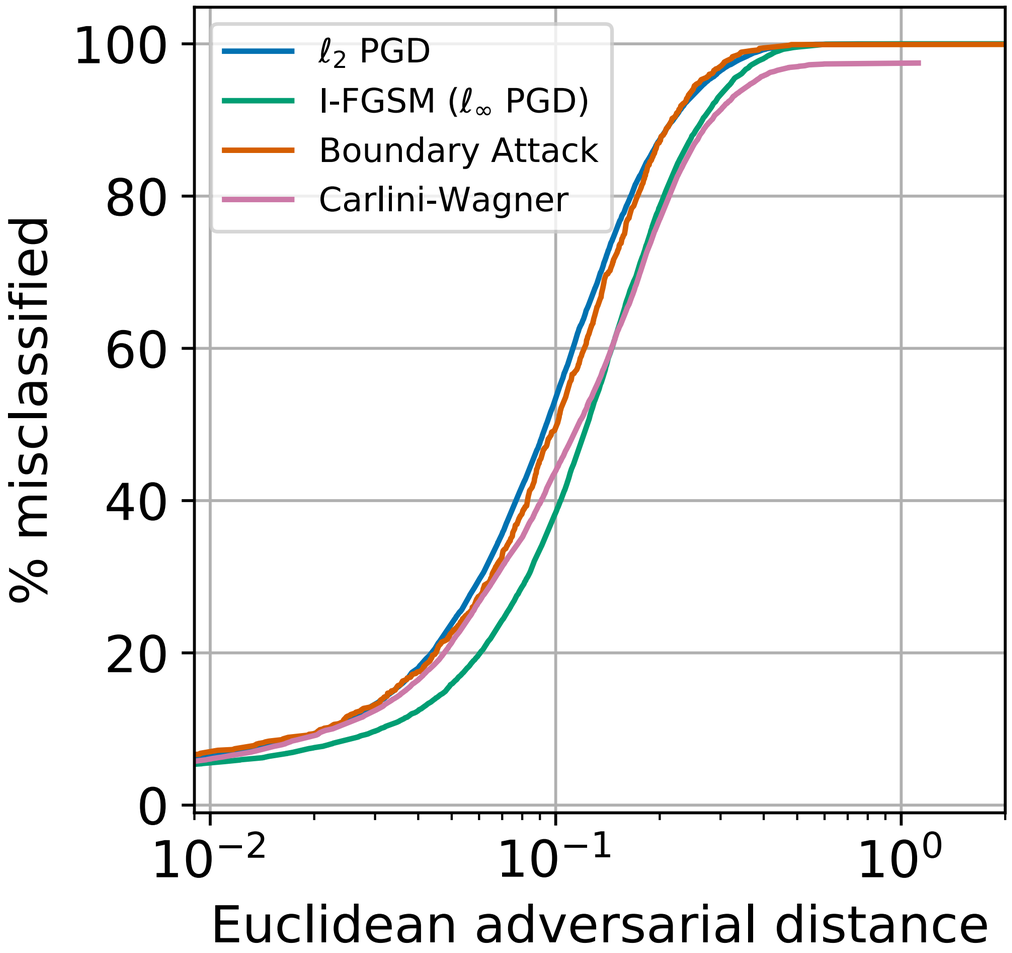
Adversarial Robustness without loss of accuracy.

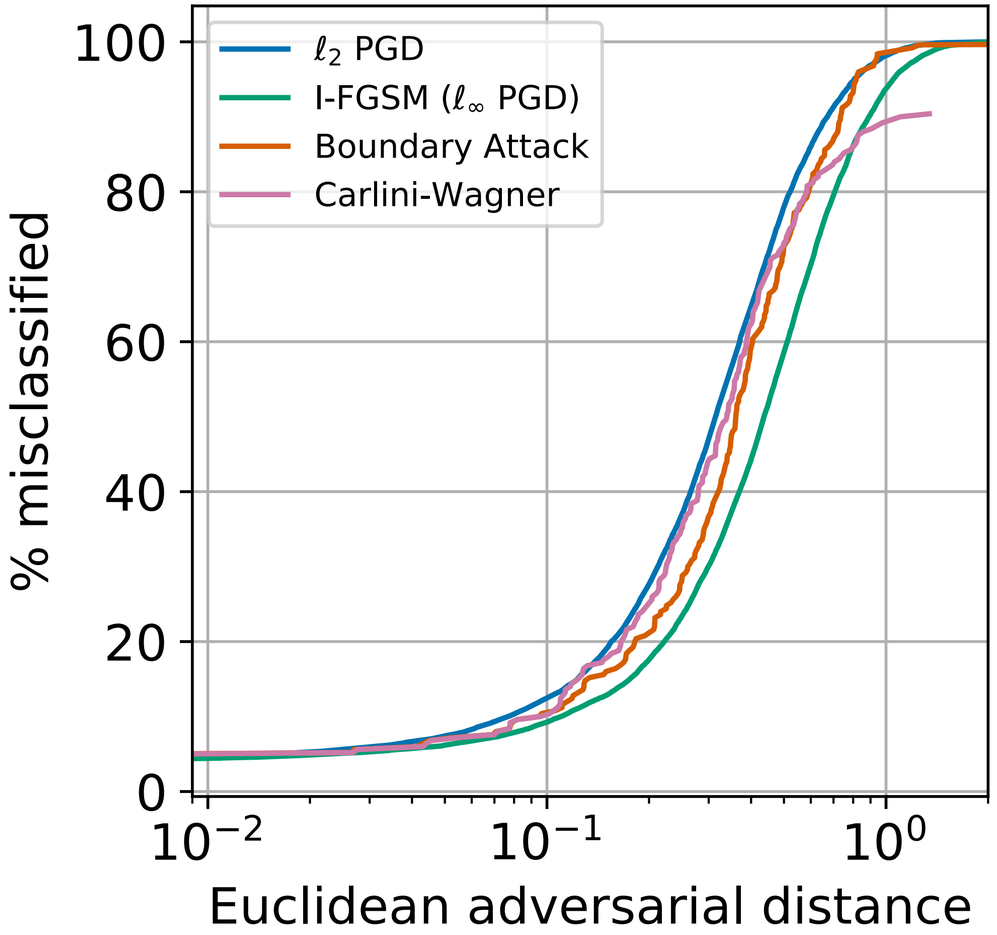
It is somewhat less robust than Madry/Qian. But no loss of accuracy on clean test images. Increasing lambda/epsilon gives better robustness at a slight loss of accuracy.

			Defence method		
Attack		details	ℓ_2 TV + Lipschitz $\varepsilon = 0.01$, $\lambda = 0.1$	Madry	Qian
test error		undefended model	**4.1**	4.8	5.0
		defended model	**4.1**	12.8	22.8
ℓ_2 PGD	distance	$\|\varepsilon\|_2 = 100/255$	59.8	> 90	-
CW	distance	$\|\varepsilon\|_2 = 1.5$	90.8	-	**79.6**
I-FGSM	distance	$\|\varepsilon\|_\infty = 8/255$	98.1	**54.2**	-

Defence methods on CIFAR-10. Classification error (Smaller is better). Each row corresponds to an adversarial attack method (in 2-norm or infinity-norm) Results from Finlay-Abassi-Oberman 2018.

Results





Comparison of attack methods using error curves for ResNeXt-34 (2x32), on the CIFAR-10 test set. Error curve. Left: undefended model; right: best regularized model.

Attack Detection: Background

Carlini-Wagner Evasion Attacks

- 8 papers published in 2017 conferences, detecting attacks
- Detection methods based on statistics of the images,
  - for example, PCA analysis of attacked images
  - most of the detection methods ignored the model.
  - The Bayesian one used multiple models
- Then Carlini-Wagner came and broke every detection method.
- They used a modified loss function, which used knowledge of the detection method to optimize
  - misclassification + undetectable

Nicolas Carlini

Our attack detection: Image Vulnerability

We propose: vulnerability to attack as a proxy for attack detection.
Uses the model instead of statistics of the data.

Pessimistic: if you can be attacked, assume you will be.

Will clearly generate some false positives.

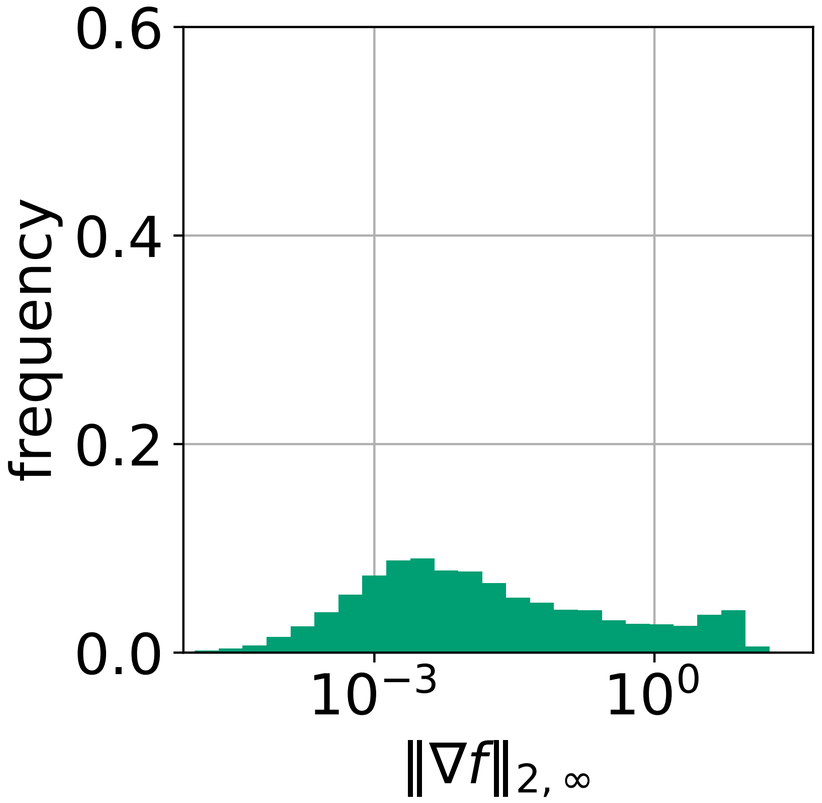
Taylor series argument (from a different part)

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon \|\nabla \ell(x)\|_* + \mathcal{O}(\varepsilon^2), \qquad d \text{ optimal} \tag{6}$$

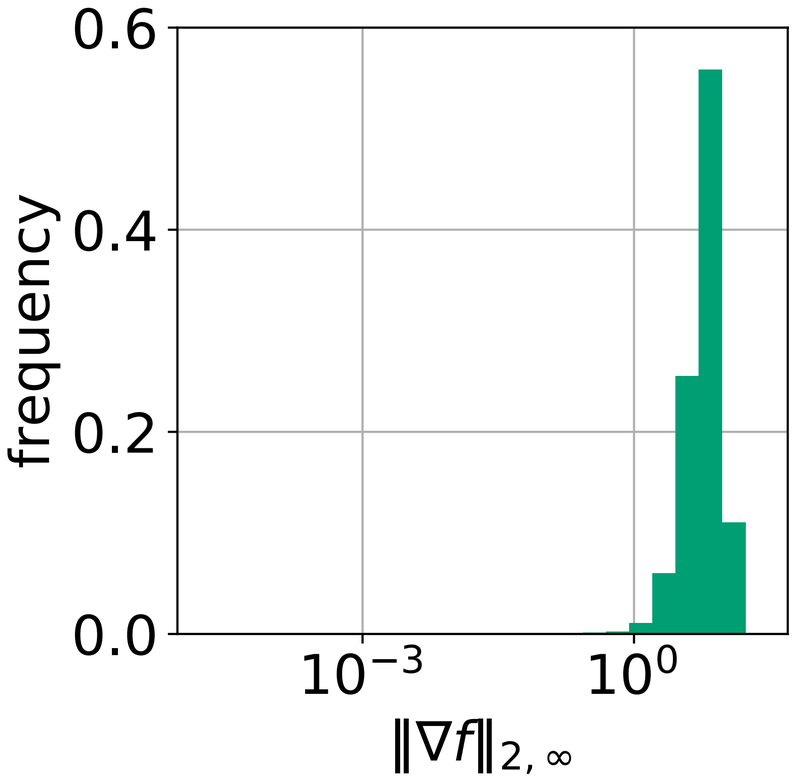
images with larger loss gradients are more vulnerable to gradient attacks.

Attack detection works?

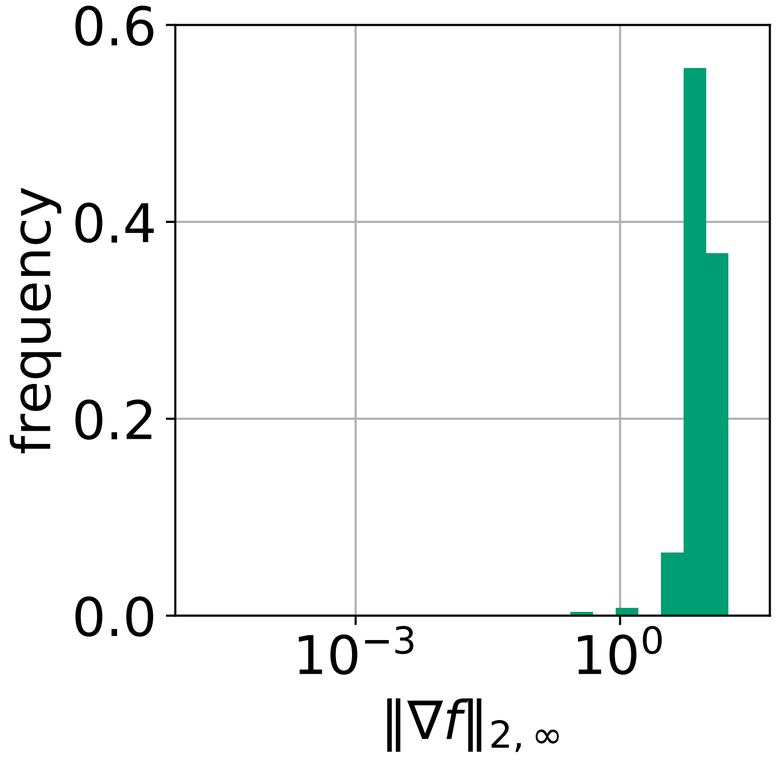
clean images have small values



middle: gradient attack



Right: boundary attack.



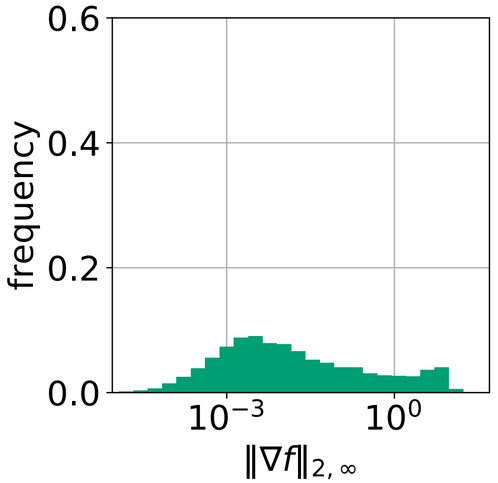
Attack detection results: All attacked images have large gradients. Most clean images have small gradients.

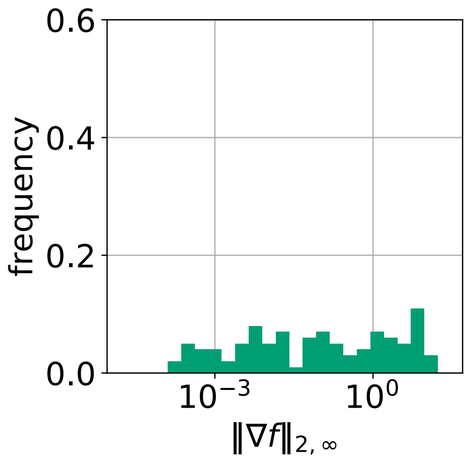
Choose a threshold so only 5% of clean images are wrongly detected attacked.

Wait: C-W style self attack.

Design a CW style modified loss attack. Now instead of image statistics, we try to attack while also making gradient small.
Results: attacks can succeed, but require much larger attack distance.

	Image source				
	clean	PGD	Boundary	CW	evasive CW
attack detected?	6%	96%	100%	100%	22%
median ℓ_2	-	0.31	0.36	0.34	0.81





Clean and evasive attack histogram

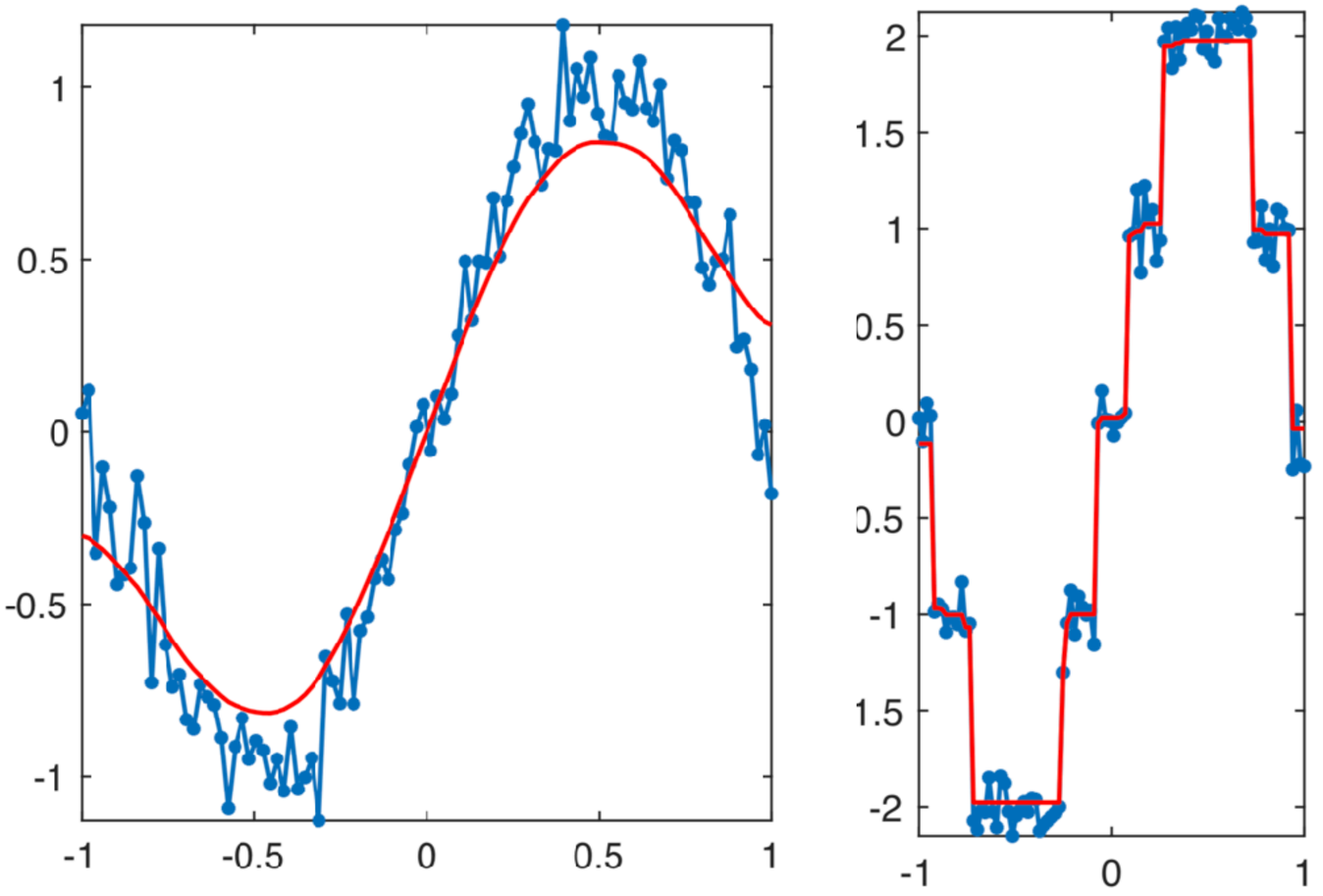
Detection Conclusion

The fact that designing evasive attacks makes the median distance so large implies that our defence method is fairly robust.
From a game theory perspective, even knowing the detection method, the attacker cannot avoid detection without making a much larger distance attack.
The attacks are not quite visible, but much closer than the detectable attacks.

Introduction to Variational Regularization

Loss depends on noise - regularizer depends on signal

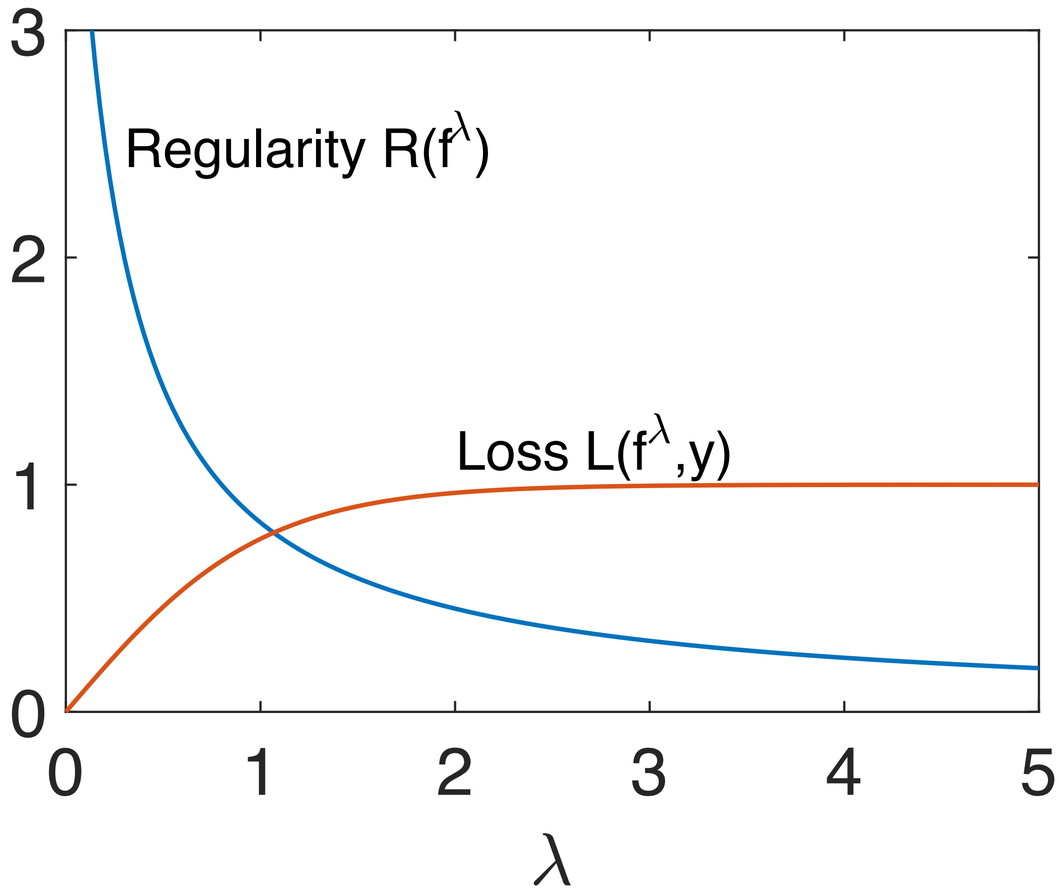
- In Machine Learning, we choose a loss function which is designed with the statistics of the noise in mind.
  - E.g. quadratic loss for Gaussian noise
- In Calculus of Variations, we choose a regularizer based on the smoothness properties of the model/function.
  - Tychonoff regularization for smooth (left)
  - Total Variation regularization for piecewise smooth (right)



Noisy and Regularized signal

Mathematics: Calculus of Variations

- Derive physical laws from energy minimization principles.
- “Path of least resistance” - Variational principle for the path of light in a medium
- Often there is a single non-dimensional parameter that determines the “wildness” of the system. For example the Reynolds number for fluid dynamics.
- Regularization: adding “friction” to the system. Leads to smoother solutions.



Regularity and Loss as a function of smoothing parameter

Total Variation Denoising [1992] R-Osher-F.

used in early, high profile image reconstruction of video images.

$$J[u] = L[u; u_0] + \lambda R[\nabla u]$$

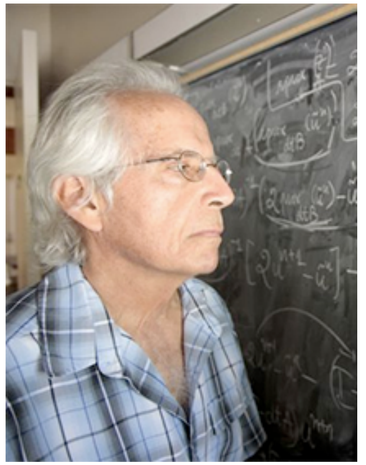
Original

Noisy image

Denoised image

- minimize a variational functional: combination of a loss term, to the original noisy image, and a regularization term

- Regularization is large on noise, small on images.
  Regularization: **Total Variation.**



Stanley Osher

Image inpainting [B. Sapiro, Casselles, B. ’00]

Fill in missing parts of image, without adding additional information. *Analogy with generalization*

$$J[u] = L[u; u_0] + \lambda R[\nabla u]$$

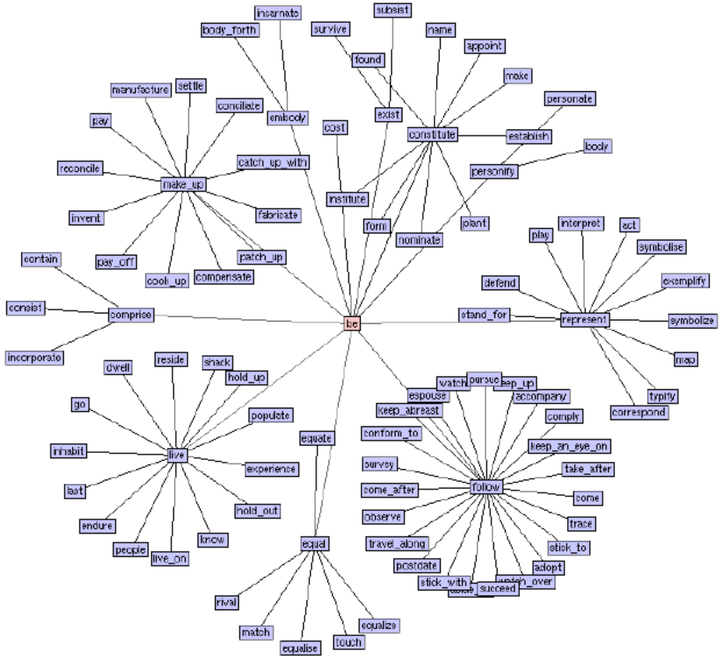
- minimize a variational functional: combination of a loss term to the given image (on the data manifold), and a regularization term
- Regularization: **Lipschitz Regularization**
- Equivalent to solving the Infinity-Laplace PDE [O. ’04, ’13]
- FYI [Peres Tug-of-War and IL ’09]

G. Sapiro

Regularization of models

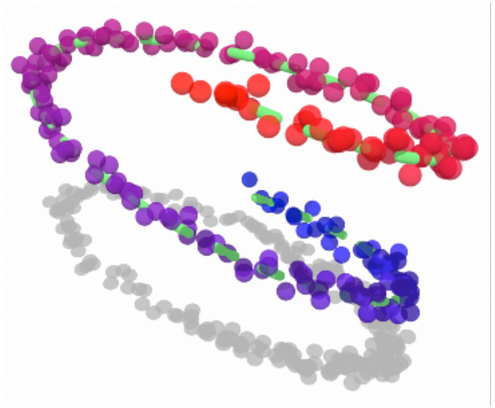
Idea: apply a regularizers to the model; the function mapping data to labels.

f(x)

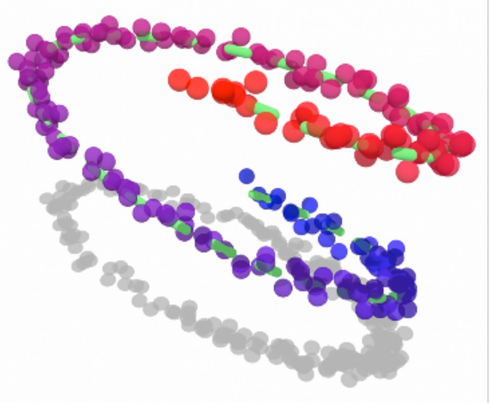


x in M = manifold of images

word labels



Q: Can we find and implement a good regularizer which promotes adversarial robustness? (A: yes)



unregularized map: well-behaved on data manifold, but very bad off the manifold (without regularization)

Interpretation: regularization fixes large gradients (instability to perturbations) on, or near, the data manifold

Adversarial Robustness measures : vulnerability of a model to adversarial attack.

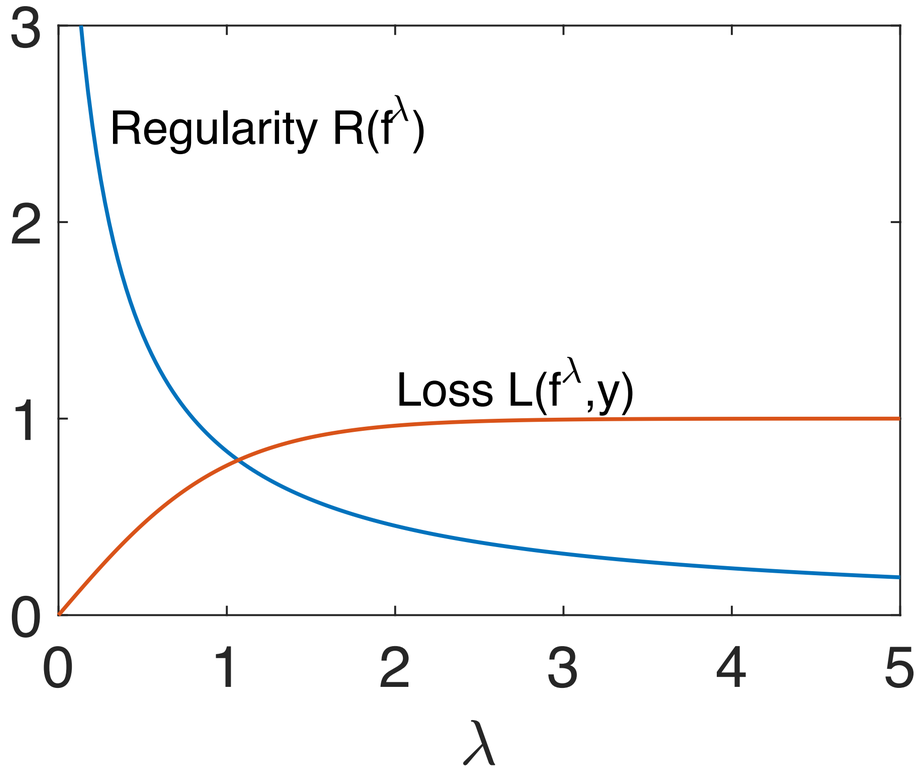
Weng et al. (2018) and Hein & Andriushchenko (2017) propose the the Lipschitz constant of the model.
So try Lipschitz Regularization?

We will show that a modified, more accurate version of Adversarial Training corresponds to Total Variation Regularization. Best results come from combining both.

Outcome of the regularized model

Our Regularization:
lambda Total Variation + epsilon Lipschitz

Attack	details	Defence method: ℓ_2 TV + Lipschitz, $\varepsilon = 0.01$, $\lambda = 0.1$	Defence method: ℓ_2 TV + Lipschitz, $\varepsilon = 0.1$, $\lambda = 1$
test error	undefended model	**4.1**	**4.1**
	defended model	**4.1**	6.0
ℓ_2 PGD	$\|\varepsilon\|_2 = 100/255$	59.8	**36.1**
CW	distance $\|\varepsilon\|_2 = 1.5$	90.8	84.4
I-FGSM	$\|\varepsilon\|_\infty = 8/255$	98.1	93.7



Defence methods on CIFAR-10. Classification error (percentage) on test images.
Increasing regularization parameters by factor of 10 leads to better robustness, but loss of accuracy. Consistent with cartoon.

Adversarial attacks as an optimization problem

Adversarial Attacks on the loss

Write $y = f(x, w)$ for the model with input x and parameters w
Let $\mathcal{L}(x, y)$ be the loss
Write $\ell(x) = \mathcal{L}(f(x, w), y)$ for the model loss.

The optimal loss attack of norm ε on the image vector, x is the solution of

(1)
$$\max_{\|x' - x\| \leq \varepsilon} \ell(x').$$

Blackboard derivation: Signed Gradient Attack Vector

The Fast Gradient Sign Method (FGSM) Goodfellow et al. (2014) arises when attacks are measured in the ∞-norm. It corresponds to a one step attack in the direction d_1 given by the signed gradient

$$(d_1)_i = \frac{\nabla \ell(x)_i}{|\nabla \ell(x)_i|}. \tag{2}$$

The attack direction (2) arises from linearization of the objective in (1), which leads to

$$\max_{\|d\|_\infty \leq 1} d \cdot \nabla \ell(x) \tag{3}$$

By inspection, the minimizer is the signed gradient vector, (2), and the optimal value is $\|\nabla \ell(x)\|_1$. When the 2-norm is used in (3), the optimal value is $\|\nabla \ell(x)\|_2$ and the optimal direction is the normalized gradient

$$\frac{\nabla \ell(x)}{\|\nabla \ell(x)\|_2}. \tag{4}$$

More generally, when a generic norm is used in (3), the maximum of the linearized objective is the dual norm (Boyd & Vandenberghe, 2004, A.1.6).

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon d \cdot \nabla \ell(x) + \mathcal{O}(\varepsilon^2)$$

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon \|\nabla \ell(x)\|_* + \mathcal{O}(\varepsilon^2), \qquad d \text{ optimal}$$

Dual norms and attacks

2.1. **Derivation of attack directions.** The solution of (3) can be approximated using the dual norm (Boyd & Vandenberghe, 2004, A.1.6). If the ∞-norm is used, we recover the Signed Gradient (Goodfellow et al., 2014). However a different attack vector is obtained if we measure attacks in the 2-norm.

Theorem 2.2. *The optimal attack vector defined by* (3) *in a generic norm* $\|\cdot\|$ *can be approximated to* $\mathcal{O}(\varepsilon^2)$ *with the vector* εa*, where* a *is the solution of*

$$a \cdot v = \|v\|_*, \qquad \text{with } v = \nabla_x \ell(f(x), y) \tag{4}$$

and $\|\cdot\|_*$ *is the dual norm. In particular* a *is given by*

$$\begin{cases} a_i^{SG} = \dfrac{\nabla \ell(x)_i}{|\nabla \ell(x)_i|} & \text{for the } \infty\text{-norm} \\ a^{\ell_2} = \dfrac{\nabla \ell(x)}{\|\nabla \ell(x)\|_2} & \text{for the } 2\text{-norm} \end{cases} \tag{5}$$

Blackboard: Why random attacks are weak, adversarial attacks are strong

- Blackboard/Exercise: taylor series, random attacks, mean zero, versus gradient attack

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon d \cdot \nabla \ell(x) + \mathcal{O}(\varepsilon^2)$$

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon \|\nabla \ell(x)\|_* + \mathcal{O}(\varepsilon^2), \qquad d \text{ optimal}$$

Adversarial Regularization

Derivation: Total Variation Regularizaton from Adversarial Training

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon \|\nabla \ell(x)\|_* + \mathcal{O}(\varepsilon^2)$$

The equation above shows that perturbing an image by an optimal one step attack vector is equivalent to modifying the loss with an extra term.

Take expectations (drop the higher order term)

$$\mathbb{E}\left[\ell\left(x\right)\right] + \varepsilon \underbrace{\mathbb{E}\left[\|\nabla_x \ell(x)\|_*\right]}_{\text{Total Variation}}$$

Thus adversarial training corresponds to Total Variation Regularization

Discussion: Robustness and Lipschitz constant

In deep learning, the Lipschitz constant of a model appears in the context of model robustness Xu & Mannor (2012), generalization Bartlett (1996); Sokolic et al. (2017), and Wasserstein GANs Gulrajani et al. (2017); Miyato et al. (2018); Anil et al. (2018).

In the study of adversarial robustness, Weng et al. (2018) and Hein & Andriushchenko (2017) showed that the Lipschitz constant of the model gives a certifiable minimum adversarial distance under which the model is robust to perturbations. Thus, if the Lipschitz of the model can be controlled, the model will be robust. Indeed, training models to have small Lipschitz constant has empirically been shown to improve adversarial robustness Tsuzuku et al. (2018); Cissé et al. (2017). By its very definition, the Lipschitz constant determines model robustness (sensitivity to changes in the data): The Lipschitz constant of a function f is the best constant L so that

$$\|f(x+v) - f(x)\| \leq L\|v\| \tag{11}$$

Discussion: Estimating Model Lipschitz constant

However estimating the Lipschitz constant of a deep model can be challenging. Data independent upper bounds on the Lipschitz constant of the model go back to Bartlett (1996). These bounds are based on the product of the norm of weight matrices, but the gap in the bound can grow exponentially in the number of layers, since it neglects the effects of the activation function. Alternative methods, as in Weng et al. (2018) can be costly.

We make the observation that the Lipschitz constant of a network may be estimated using *Rademacher's Theorem* (Evans, 2018, §3.1): if a model $f(x)$ is Lipschitz continuous on a (compact) set $\mathcal{D}$, then it is differentiable almost everywhere and the Lipschitz constant is given by

$$\max_{x \in \mathcal{D}} \|\nabla f(x)\|. \tag{12}$$

Thus the Lipschitz constant of a model[1] (or alternately, the loss) may be underestimated by simply choosing the maximum gradient norm over the data.

Derivation: Lipschitz Regularization

$$\ell(x + \varepsilon d) = \ell(x) + \varepsilon \|\nabla \ell(x)\|_* + \mathcal{O}(\varepsilon^2)$$

Instead of averaging, consider a penalty for the largest gradient norm

$$\cdot \lambda \underbrace{\max \|\nabla_x \ell(x)\|_*}_{\text{Lipschitz regularization}} .$$

Then by Rademacher's Theorem, this corresponds to Lipschitz Regularization

Combined Model: TV and Lip regularization

$$\mathbb{E}\left[\ell(x)\right] + \varepsilon \underbrace{\mathbb{E}\left[\|\nabla_x \ell(x)\|_*\right]}_{\text{TV regularization}} + \lambda \underbrace{\max \|\nabla_x \ell(x)\|_*}_{\text{Lipschitz regularization}} .$$

Proof of Convergence and Generalization for Lipschitz Regularized DNNs

joint with Jeff Calder

Lipschitz regularized Deep Neural Networks converge and generalize O. and Jeff Calder; 2018

Lipschitz Regularization of DNNs

Train with the expected loss augmented by the Lipschitz regularization term

$$(1) \qquad \min_{f:X\to Y} J^n[f] = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;w), u_0(x_i)) + \lambda \max(\mathrm{Lip}(f) - L_0, 0)$$

- u0 - the true label function
- L_0 - the Lipschitz constant of u0, estimated from data.

TABLE 2. Lipschitz constants of common training sets. CIFAR-100 has several duplicated images with different labels, these were removed from the calculation.

Dataset	MNIST	FashionMNIST	CIFAR-10	CIFAR-100
$\mathrm{Lip}_{2,\infty}(\mathcal{D})$	0.417	0.626	0.364	1.245

Take the limit as we sample more points.
The limiting functional is given by

$$(5) \qquad J^{Lip,\rho}[u] \equiv \int_X \ell(u(x), u_0(x)) d\rho(x) + \lambda \max(\mathrm{Lip}(u) - L_0, 0)$$

Statement of convergence theorem for Noisy Labels

$$J^{Lip,\rho}[u] \equiv \int_X \ell(u(x), u_0(x)) d\rho(x) + \lambda \max(\mathrm{Lip}(u) - L_0, 0) \tag{5}$$

Theorem 2.11. *Suppose that* $\inf_{\mathcal{M}} \rho > 0$, $\ell : Y \times Y \to \mathbb{R}$ *is Lipschitz, and let* $u^* \in W^{1,\infty}(X;Y)$ *be any minimizer of the limiting functional* (6). *Then with probability one*

$$u_n \longrightarrow u^* \quad \textit{uniformly on } \mathcal{M} = \mathrm{supp}(\rho) \textit{ as } n \to \infty,$$

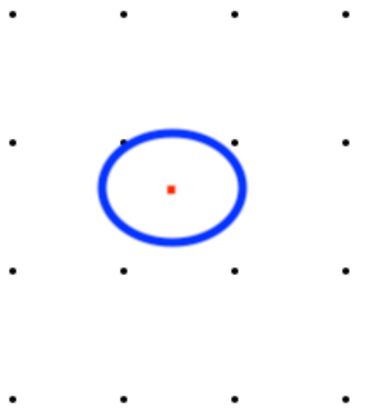
where u_n *is any sequence of minimizers of* (1). *Furthermore, every uniformly convergent subsequence of* u_n *converges on* X *to a minimizer of* (6).

Estimates of worst case distance from a point to sampled points

- n - the number of data points sampled
- m - the dimension of the data manifold.

Uniform sampling (grid)

$$\|\mathrm{Id} - \sigma_n\|_{L^\infty(\mathcal{M};X)} \leq C\left(\frac{1}{n}\right)^{1/m}$$

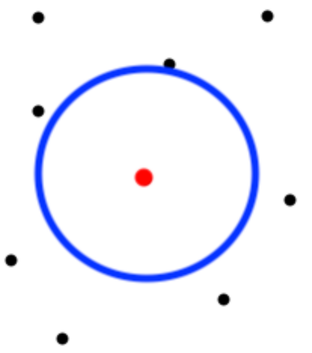


Random sampling from density

Lemma 2.9. *Suppose that* $\inf_{\mathcal{M}} \rho > 0$. *Then for any* $t > 0$

$$\|Id - \sigma_n\|_{L^\infty(\mathcal{M};X)} \leq C\left(\frac{t\log(n)}{n}\right)^{1/m}$$

with probability at least $1 - Ct^{-1}n^{-(ct-1)}$.



Convergence with a rate

$$(1) \qquad \min_{f:X\to Y} J^n[f] = \frac{1}{n}\sum_{i=1}^{n} \ell(f(x_i;w), u_0(x_i)) + \lambda \max(\mathrm{Lip}(f) - L_0, 0)$$

Theorem 2.7. *Suppose that* $\mathrm{Lip}[u_0] \le L_0$ *and* $\inf_{x\in\mathcal{M}} \rho(x) > 0$. *If* f_n *is any sequence of minimizers of* (1) *then for any* $t > 0$

$$\|u_0 - f_n\|_{L^\infty(\mathcal{M};Y)} \le C L_0 \left(\frac{t\log(n)}{n}\right)^{1/m}$$

holds with probability at least $1 - Ct^{-1}n^{-(ct-1)}$.

- u0 - the true label function
- L_0 - the Lipschitz constant of u0, easily estimated from data.
- *n* - the number of data points sampled
- *m* - the dimension of the data manifold.

Generalization follows

As an immediate corollary, we can prove that the generalization loss converges to zero, and so we obtain perfect generalization.

Corollary 2.8. *Assume that for some $q \geq 1$ the loss ℓ satisfies*

$$\ell(y, y_0) \leq C\|y - y_0\|_Y^q \quad \text{for all } y_0, y \in Y. \tag{6}$$

Then under the assumptions of Theorem 2.7

$$L[u_n, \rho] \leq C L_0^q \left(\frac{t \log(n)}{n}\right)^{q/m}$$

holds with probability at least $1 - Ct^{-1}n^{-(ct-1)}$.

Proof. By (6), we can bound the generalization loss as follows

$$L[u_n, \rho] = \int_{\mathcal{M}} \ell(u_n(x), u_0(x))\, dVol(x) \leq CVol(\mathcal{M})\|u_n - u_0\|_{L^\infty(\mathcal{M};Y)}^q.$$

The proof is completed by invoking Theorem 2.7. □

Question: can we get a better rate for generalization with a stronger estimate?

End