First-order Adversarial Vulnerability of Neural Networks and Input Dimension

C.-J. Simon-Gabriel †, Y. Ollivier ‡, L. Bottou ‡, B. Schölkopf †, D. Lopez-Paz ‡

† Max-Planck-Institute for Intelligent Systems
‡ Facebook AI Research

Relation to Literature

Relation to Literature

[1,2] : “Under specific data assumptions, vulnerability Increases with input dimension.”

[1] Adversarial Spheres, Gilmer et al., ICLR Workshop 2018
[2] Are adversarial examples inevitable?, Shafahi et al., ICLR 2019

Relation to Literature

[1,2] : “Under specific data assumptions, vulnerability Increases with input dimension.”

- No-free-lunch-like result:
  “If data can be anything, then there exists datasets that make the problem arbitrarily hard”

[1] Adversarial Spheres, Gilmer et al., ICLR Workshop 2018
[2] Are adversarial examples inevitable?, Shafahi et al., ICLR 2019

Relation to Literature

[1,2] : "Under specific data assumptions, vulnerability Increases with input dimension."

- No-free-lunch-like result:
  "If data can be anything, then there exists datasets that make the problem arbitrarily hard"

- Cannot apply to image-datasets, because humans are a non vulnerable classifiers for which higher dimension (higher resolution) *helps*.

[1] Adversarial Spheres, Gilmer et al., ICLR Workshop 2018
[2] Are adversarial examples inevitable?, Shafahi et al., ICLR 2019

Relation to Literature

[1,2] : "Under specific data assumptions, vulnerability Increases with input dimension."

- No-free-lunch-like result:
  "If data can be anything, then there exists datasets that make the problem arbitrarily hard"

- Cannot apply to image-datasets, because humans are a non vulnerable classifiers for which higher dimension (higher resolution) *helps*.

- Hence the question:
  not : what's wrong with our data?
  but : what's wrong with our classifiers?

[1] Adversarial Spheres, Gilmer et al., ICLR Workshop 2018
[2] Are adversarial examples inevitable?, Shafahi et al., ICLR 2019

Relation to Literature

[1,2] : "Under specific data assumptions, vulnerability Increases with input dimension."

Here : "Under specific classifier assumptions, vulnerability Increases with input dimension."

- No-free-lunch-like result:
  "If data can be anything, then there exists datasets that make the problem arbitrarily hard"

- Cannot apply to image-datasets, because humans are a non vulnerable classifiers for which higher dimension (higher resolution) *helps*.

- Hence the question:
  not : what's wrong with our data?
  but : what's wrong with our classifiers?

[1] Adversarial Spheres, Gilmer et al., ICLR Workshop 2018
[2] Are adversarial examples inevitable?, Shafahi et al., ICLR 2019

Main Theorem

Main Theorem

Theorem:

At initialization, using "He-initialization", and for a very wide class of neural nets, adversarial damage increases like $\sqrt{d}$ (d: input dimension).

Main Theorem

Theorem:

At initialization, using “He-initialization”, and for a very wide class of neural nets, adversarial damage increases like $\sqrt{d}$ (d: input dimension).

Remarks:

- Vulnerability is independent of the network topology (inside a wide class).

Main Theorem

Theorem:

At initialization, using "He-initialization", and for a very wide class of neural nets, adversarial damage increases like $\sqrt{d}$ (d: input dimension).

Remarks:

- Vulnerability is independent of the network topology (inside a wide class).
- Includes any succession of FC-, conv-, ReLU-, and subsampling layers at He-init.

Main Theorem

Theorem:

At initialization, using “He-initialization”, and for a very wide class of neural nets, adversarial damage increases like $\sqrt{d}$ (d: input dimension).

Remarks:

- Vulnerability is independent of the network topology (inside a wide class).
- Includes any succession of FC-, conv-, ReLU-, and subsampling layers at He-init.

Question:

Does it hold after training? $\rightarrow$ Experiments

Experimental Setting

Experimental Setting

- Up-sample CIFAR-10
  - Yields 4 datasets with input sizes:
    (3x)32x32, 64x64, 128x128, 256x256.

Experimental Setting

- Up-sample CIFAR-10
  - Yields 4 datasets with input sizes:
    (3x)32x32, 64x64, 128x128, 256x256.

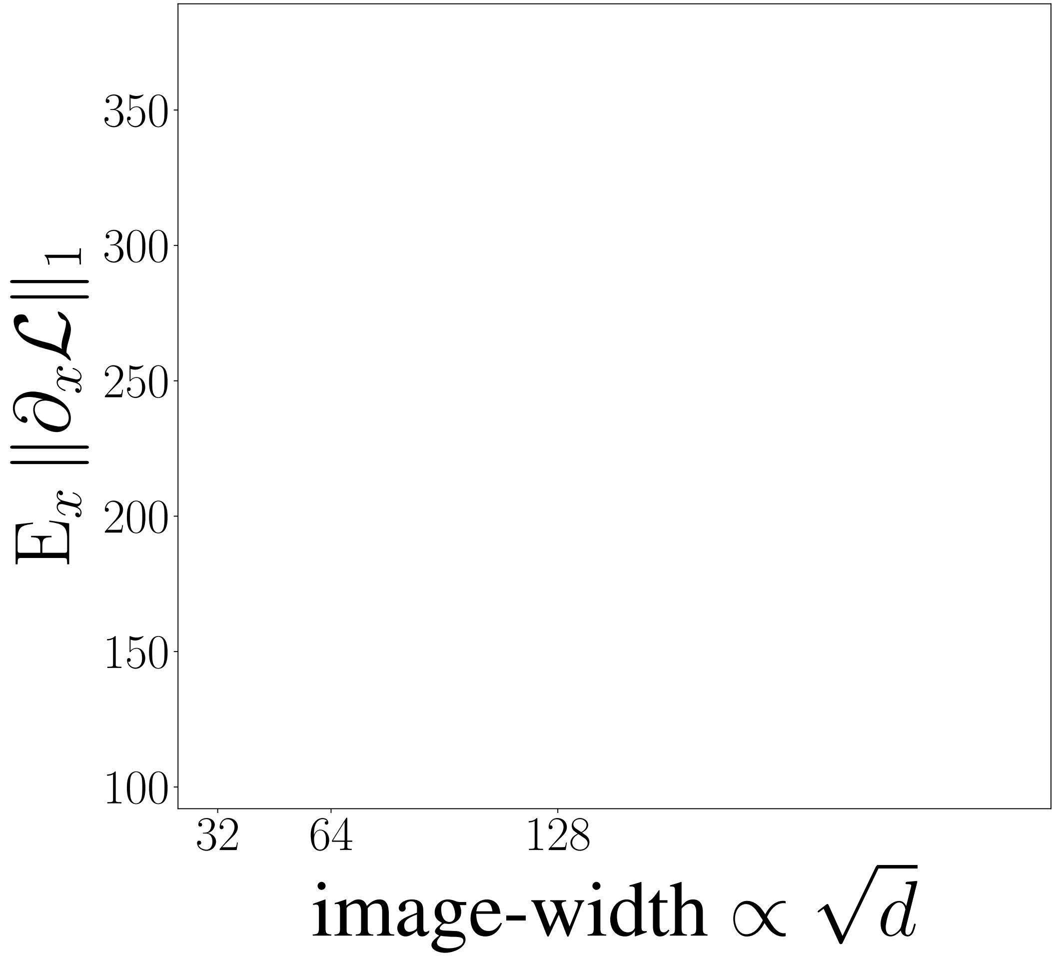
- Train a conv net for each input size
  - Use same architecture for all networks
    (up to convolution dilation and subsampling layers).

Experimental Setting

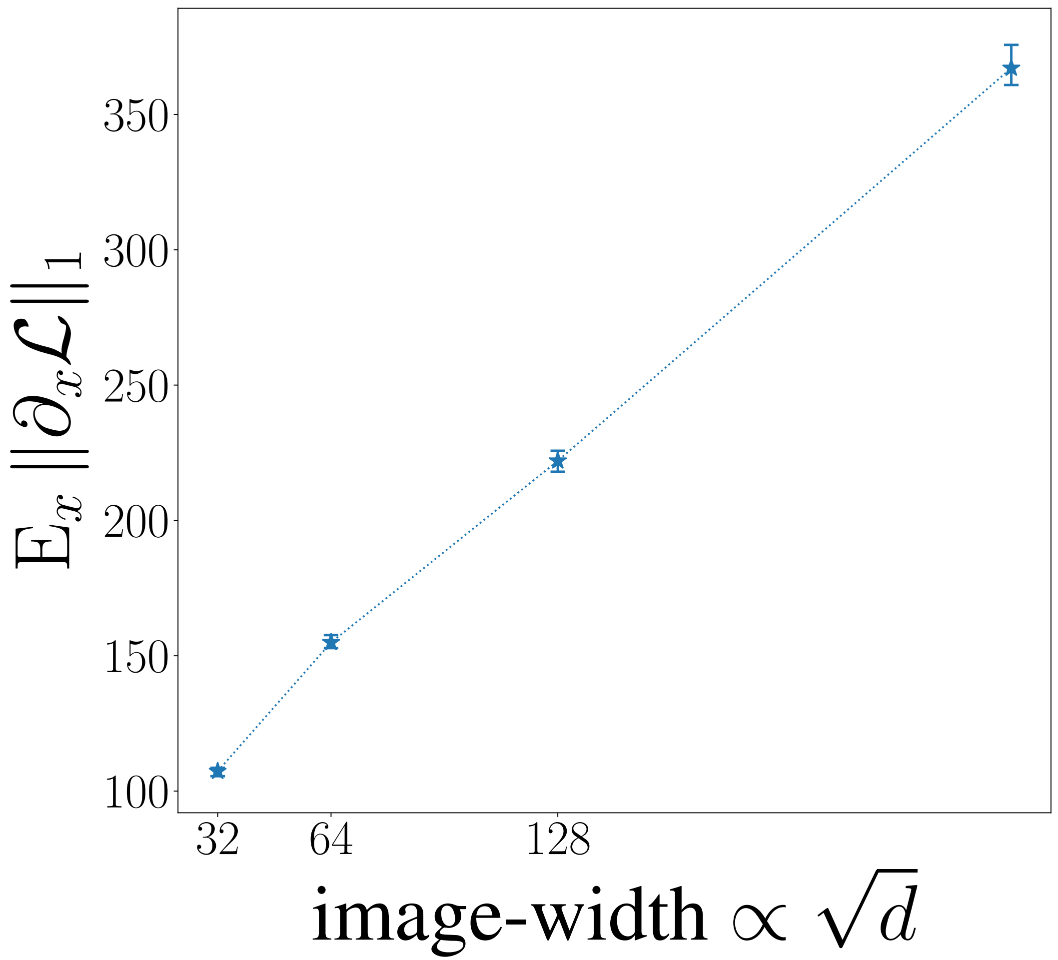
- Up-sample CIFAR-10
  - Yields 4 datasets with input sizes:
    (3x)32x32, 64x64, 128x128, 256x256.
- Train a conv net for each input size
  - Use same architecture for all networks
    (up to convolution dilation and subsampling layers).
- Compare their adversarial vulnerability

Experimental Results (after training)

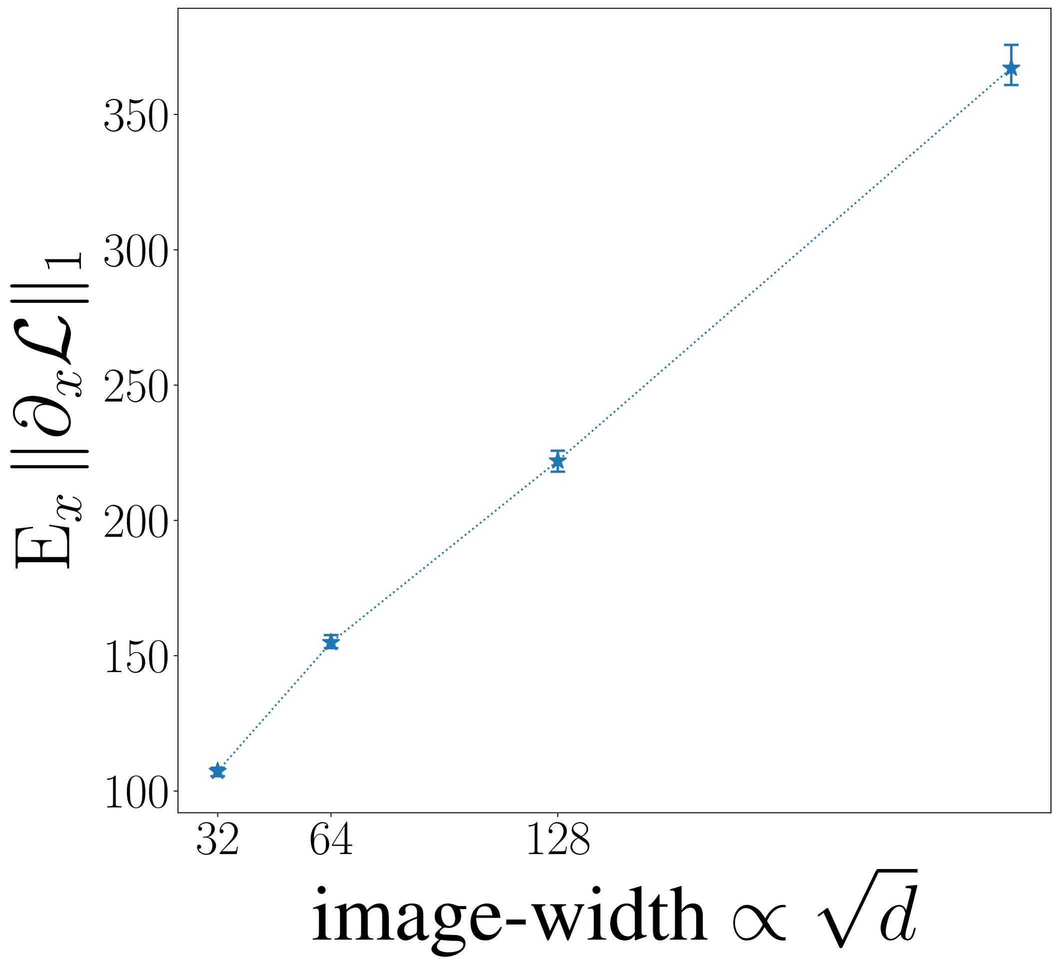
Experimental Results (after training)



Experimental Results (after training)



Experimental Results (after training)



Adversarial damage $\propto \sqrt{d}$

Conclusion

Conclusion

We show:

- Our networks are vulnerable *by design:* vulnerability increases like $\sqrt{d}$.

Conclusion

We show:

- Our networks are vulnerable *by design:* vulnerability increases like $\sqrt{d}$.
  - Proven theoretically at initialization

Conclusion

We show:

- Our networks are vulnerable *by design:* vulnerability increases like $\sqrt{d}$.
  - Proven theoretically at initialization
  - Verified empirically after usual and robust training

Conclusion

We show:

- Our networks are vulnerable *by design:* vulnerability increases like $\sqrt{d}$.
  - Proven theoretically at initialization
  - Verified empirically after usual and robust training
  - Theoretical result is independent of network topology

Conclusion

We show:

- Our networks are vulnerable *by design:* vulnerability increases like $\sqrt{d}$.
  - Proven theoretically at initialization
  - Verified empirically after usual and robust training
  - Theoretical result is independent of network topology

Suggests that:

- Current networks are not yet data-specific enough.

Conclusion

We show:

- Our networks are vulnerable *by design:* vulnerability increases like $\sqrt{d}$.
  - Proven theoretically at initialization
  - Verified empirically after usual and robust training
  - Theoretical result is independent of network topology

Suggests that:

- Current networks are not yet data-specific enough.
- Architectural tweaks may not be sufficient to solve adversarial vulnerability.

Thank you for listening!

Yann Ollivier

Léon Bottou

Bernhard Schölkopf

David Lopez-Paz

First-order Adversarial Vulnerability of Neural Networks and Input Dimension

Poster Pacific Ballroom #62