Intelligence Artificielle Avancée (RCP211)
Robustesse décisionnelle
Stabilité et généralisation

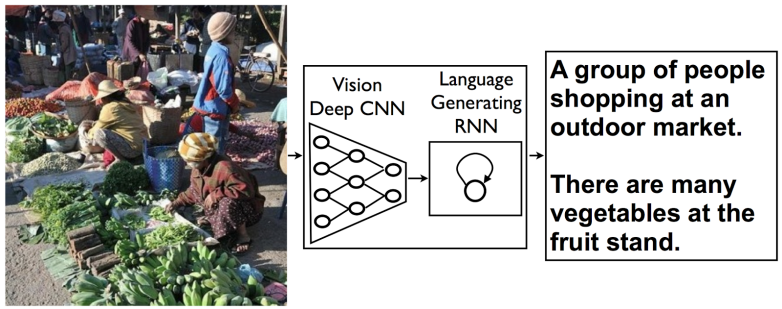
Nicolas Thome
Conservatoire National des Arts et Métiers (Cnam)
Laboratoire CEDRIC - équipe Vertigo

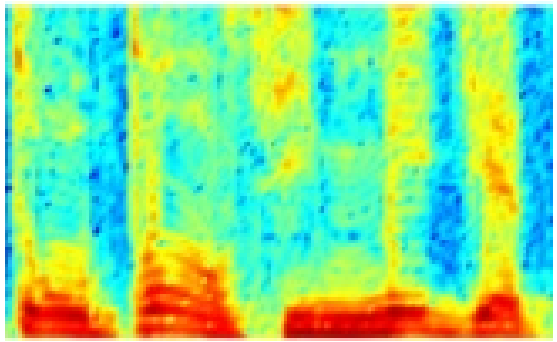
le cnam

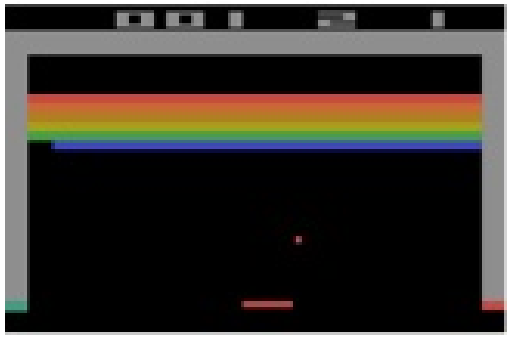
Cédric

Outline

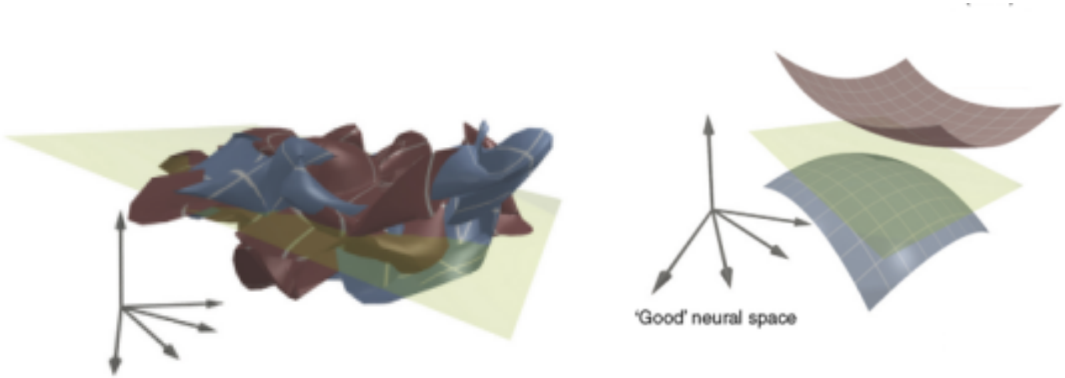
Deep Learning Theory





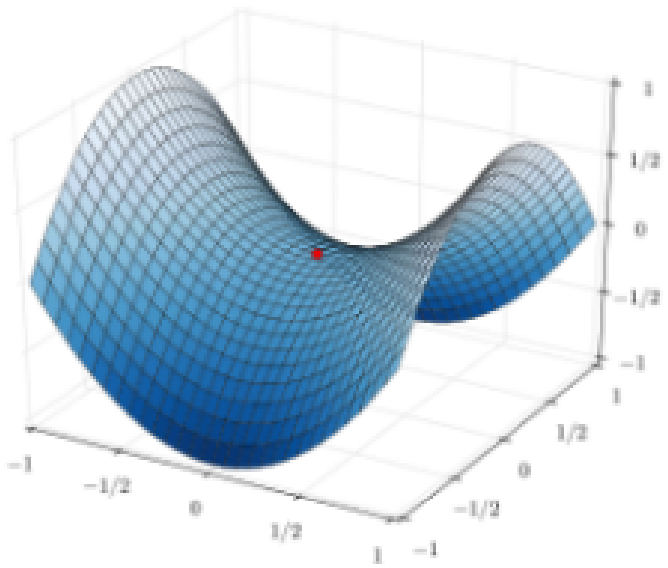


- Deep Learning: huge impact in terms of experimental results
- BUT: formal understanding still limited, other robustness issues
  - Optimization: non-convex problem
  - Generalization & over-fitting
  - Stability of the decision function
  - Explainability of the decision



Non-Convex Optimization

- One of the main historical shortcoming of deep neural networks
- In pratice, not really an issue with modern neural networks, WHY?
- Some preliminary answer elements:
  - In high dimension, few local minima
    but many saddle points [Dauphin et al., 2014]
    - Empirically, gradient descent methods manage to
      escape [Goodfellow and Vinyals, 2015] saddle points



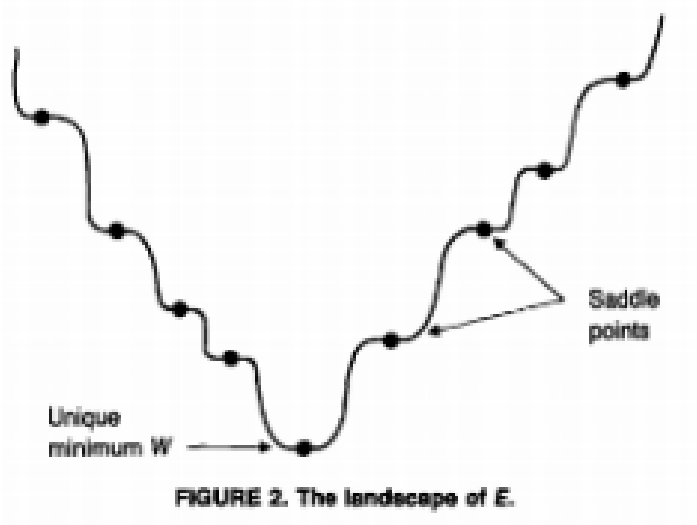
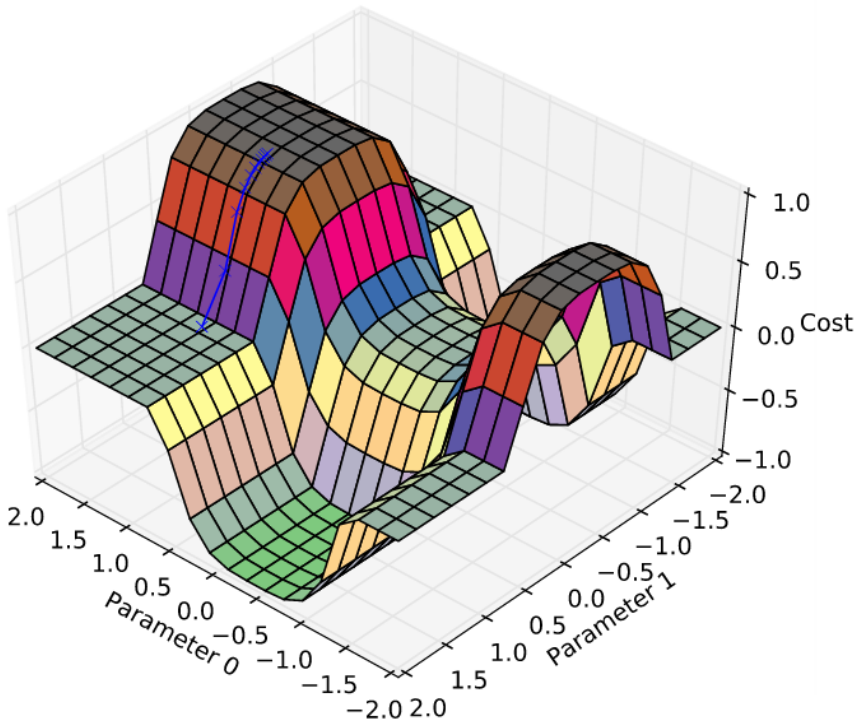


FIGURE 2. The landscape of E.

Non-Convex Optimization

- WHY non-convex optimization ist not a major practical issue for deep learning?
- Some preliminary answer elements:
  - Most of local minima have about the same objective value [Haeffele and Vidal, 2015, Choromanska et al., 2014]

(Cartoon of Dauphin et al 2014's worldview)

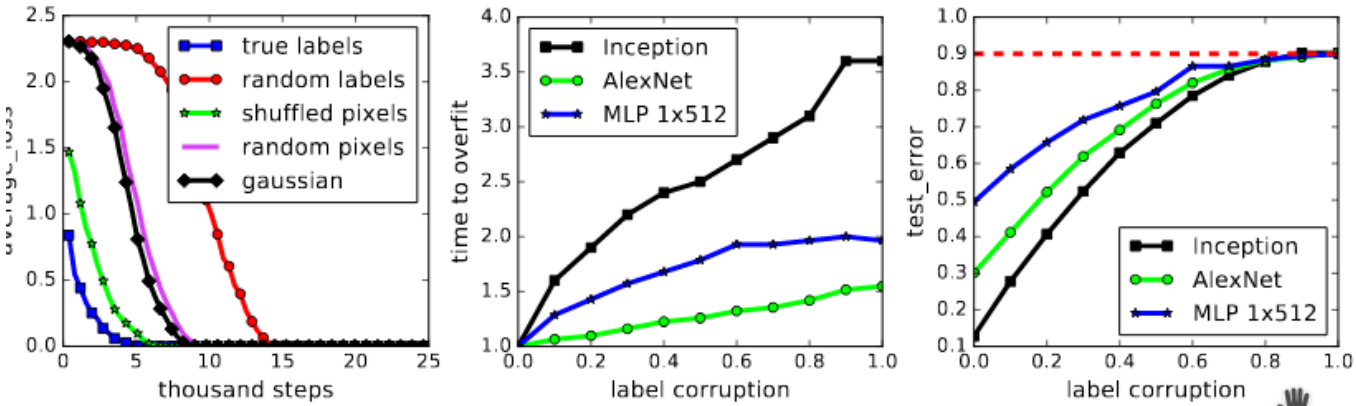


Deep Learning and generalization

- Rademacher complexity: capacity of a model to fit random label :

$$\mathcal{R}_n(\mathcal{H}) = E_\sigma \left[sup_{h \in \mathcal{H}} \frac{1}{n} \sum_{i=1}^{n} \sigma_i h(x_i) \right]$$

- Rethinking generalization: Zhang *et. al.* ICLR17 [Zhang et al., 2017]



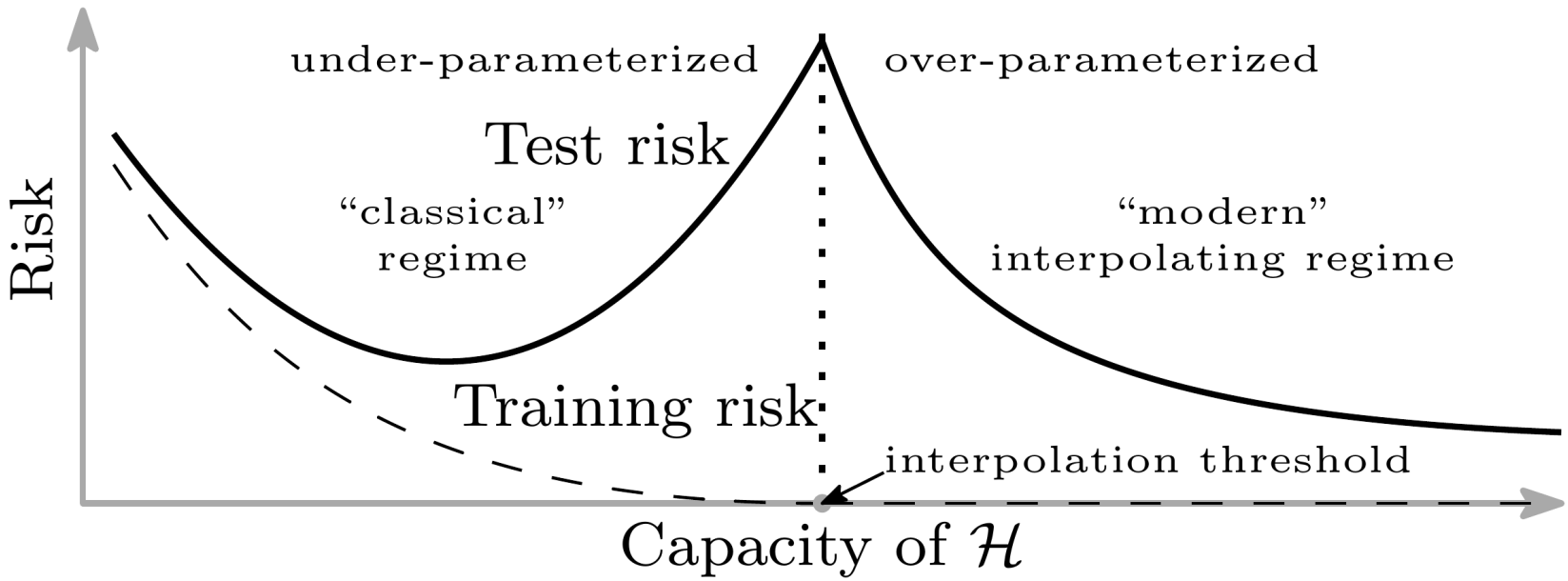
(a) learning curves (b) convergence slowdown (c) generalization error growth

  - Deep models easily fits random labels !!
  - $\mathcal{R}_n(\mathcal{H}) \approx 1 \Rightarrow$ no theoretical guarantee on generalization performances

- **Classical learning theory insufficient to explain the good generalization behavior of deep models**

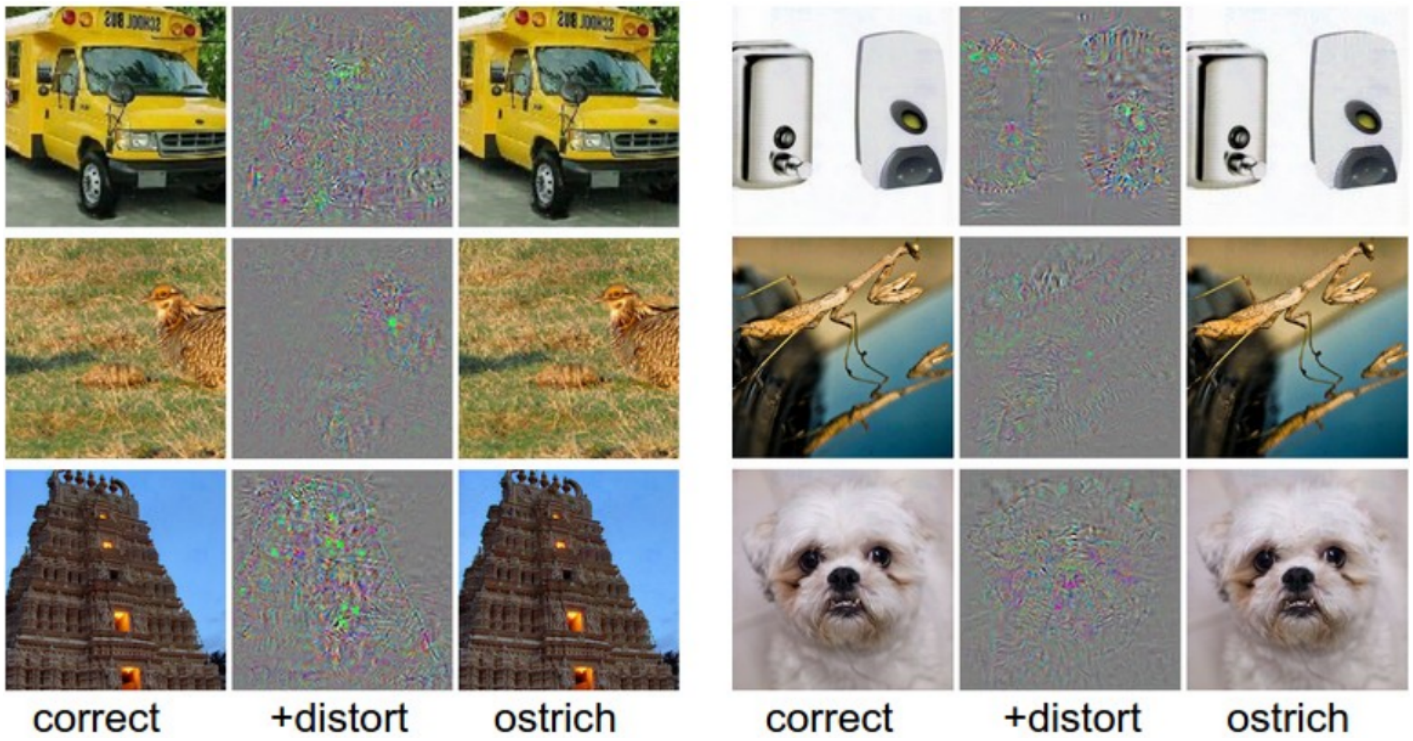
Generalization and over-parametrized models

- Double U-curve phenomena observed with deep models! [Belkin et al., 2019]



Deep Learning (DL) & Stability

- **Stability:** decision function with "controlled" variations
  - Small input variations $\Leftrightarrow$ reasonably small output variations on decision, *e.g.* Lipschitz property
  - **Decision function of deep Models not always stable**
    - **Ex: Adversarial Examples**



Deep Learning (DL) & Stability

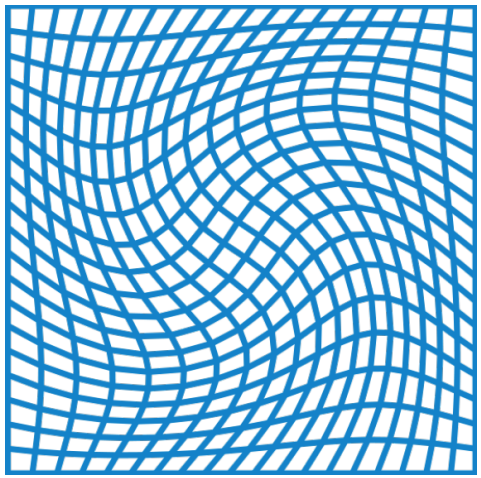
- Adversarial attacks in real-world

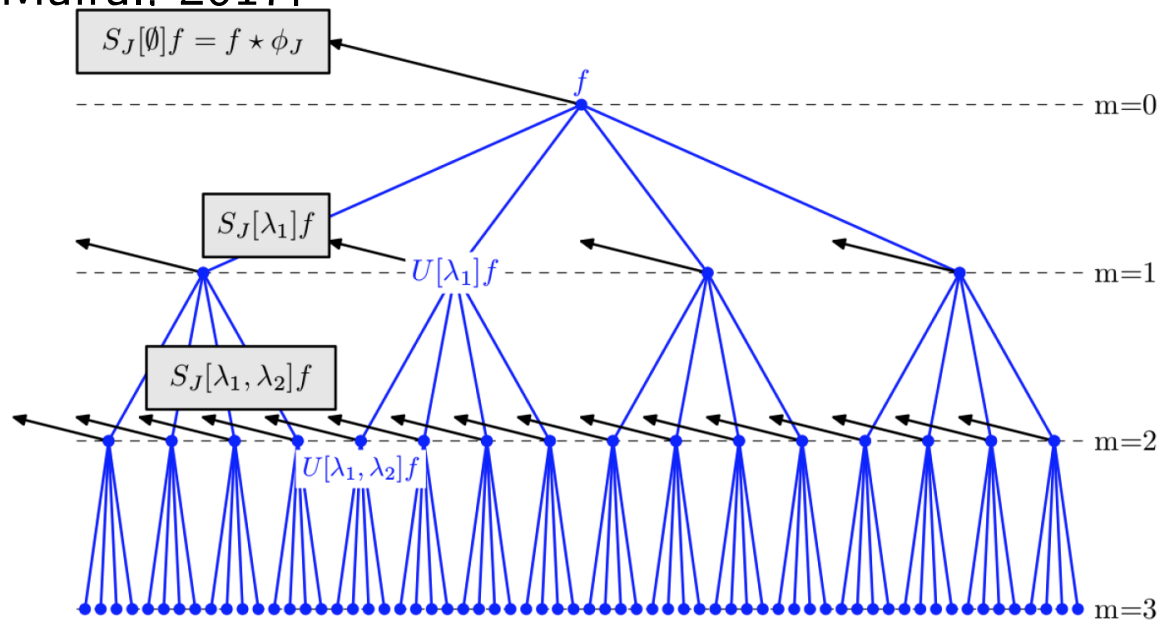
[Evtimov et al., 2017]

Deep Learning (DL) & Stability

Formal stability analysis of deep models

- Harmonic analysis in scattering operators [Mallat, 2012, Bruna and Mallat, 2013], *i.e.* "deep wavelets"
  - Show stability / invariance to diffeomorphisms
  - Stability bounds
- Generalized to deep kernel machines, closer to SoTA deep ConvNet architectures [Bietti and Mairal, 2017]

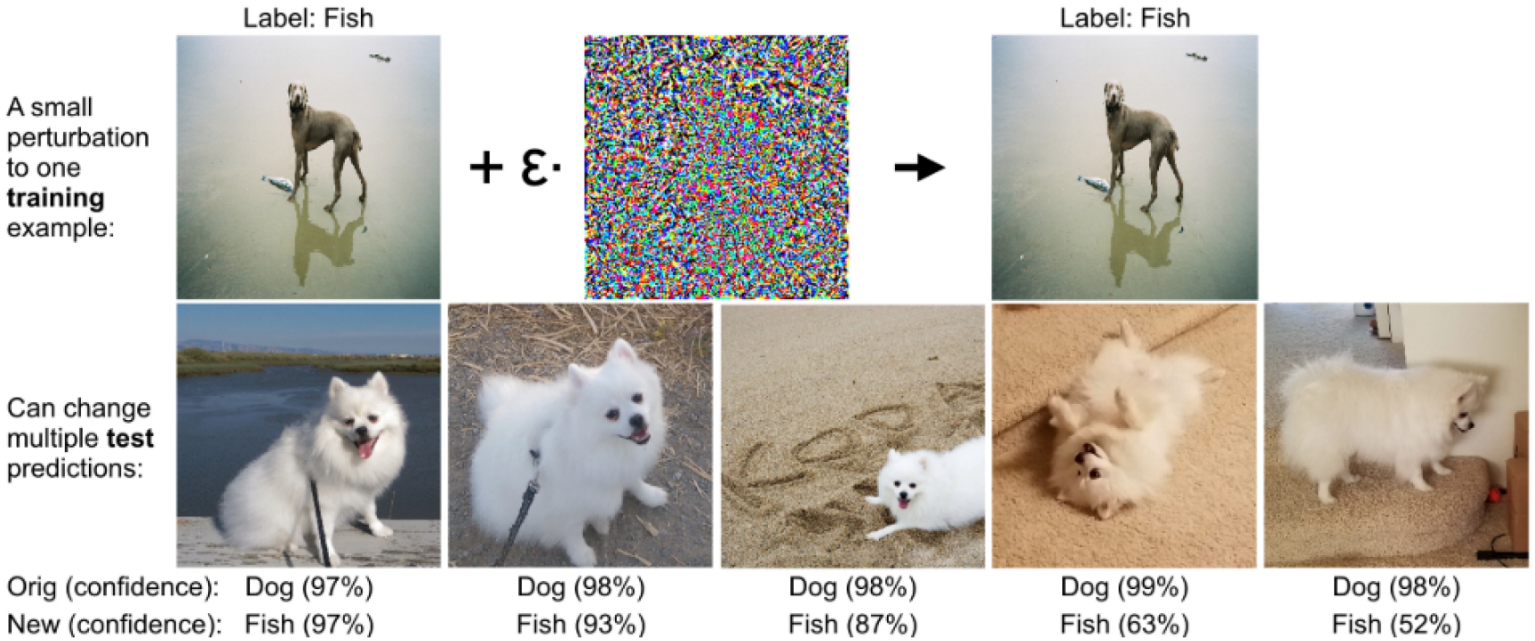




Deep Learning (DL) & Stability

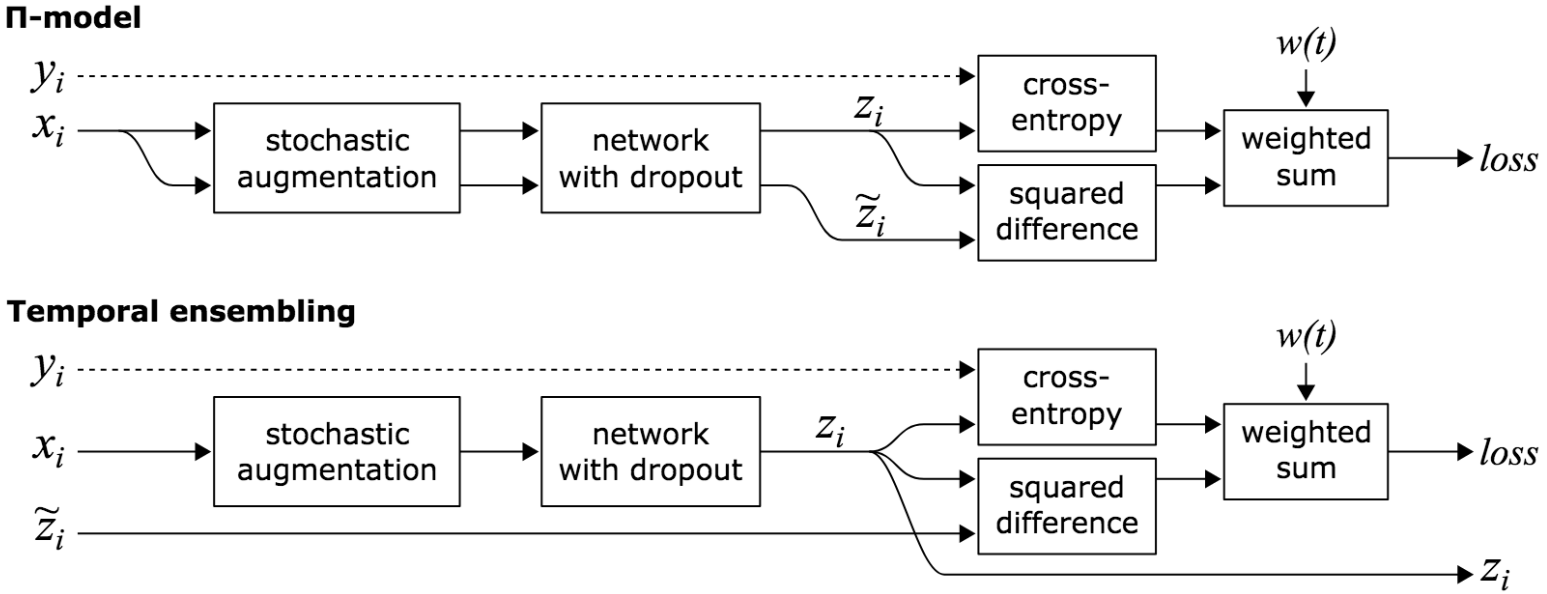
Formal stability analysis of deep models

- Influence Functions [Cook and Weisberg, 1980]
  - Characterize decision function influence on training examples
    - Removing a training point: $\mathcal{I}_{up,loss}(z, z_{test}) = -\nabla_\theta L(z_{test}, \hat{\theta})^T H_\theta^{-1} \nabla_\theta L(z, \hat{\theta})$
    - Perturbing it: $\mathcal{I}_{pert,loss}(z, z_{test})^T = -\nabla_\theta L(z_{test}, \hat{\theta})^T H_\theta^{-1} \nabla_x \nabla_\theta L(z, \hat{\theta})$
  - Adapted / applied to deep networks [Koh and Liang, 2017]
- Data poisoning



Ad hoc stability training

- Regularization criterion supporting learning stable decision function
  - Underlying model might not be stable, but helps to focus on a subet of stable functions of the family
- Robustness of the decision to transformations [Sajjadi et al., 2016], stability accross iterations [Laine and Aila. 2017. Tarvainen and Valpola. 2017]



References I

[Belkin et al., 2019] Belkin, M., Hsu, D., Ma, S., and Mandal, S. (2019).
Reconciling modern machine-learning practice and the classical bias–variance trade-off.
Proceedings of the National Academy of Sciences, 116(32):15849–15854.

[Bietti and Mairal, 2017] Bietti, A. and Mairal, J. (2017).
Invariance and stability of deep convolutional representations.
In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 6210–6220. Curran Associates, Inc.

[Bruna and Mallat, 2013] Bruna, J. and Mallat, S. (2013).
Invariant scattering convolution networks.
IEEE Trans. Pattern Anal. Mach. Intell., 35(8):1872–1886.

[Choromanska et al., 2014] Choromanska, A., Henaff, M., Mathieu, M., Arous, G. B., and LeCun, Y. (2014).
The loss surface of multilayer networks.
CoRR, abs/1412.0233.

[Cook and Weisberg, 1980] Cook, R. and Weisberg, S. (1980).
Characterizations of an empirical influence function for detecting influential cases in regression.
Technometrics, 22(4):495–508.

[Dauphin et al., 2014] Dauphin, Y., Pascanu, R., Gülçehre, Ç., Cho, K., Ganguli, S., and Bengio, Y. (2014).
Identifying and attacking the saddle point problem in high-dimensional non-convex optimization.
CoRR, abs/1406.2572.

[Evtimov et al., 2017] Evtimov, I., Eykholt, K., Fernandes, E., Kohno, T., Li, B., Prakash, A., Rahmati, A., and Song, D. (2017).
Robust physical-world attacks on machine learning models.
CoRR, abs/1707.08945.

[Goodfellow and Vinyals, 2015] Goodfellow, I. J. and Vinyals, O. (2015).
Qualitatively characterizing neural network optimization problems.
In *ICLR*.

References II

[Haeffele and Vidal, 2015] Haeffele, B. D. and Vidal, R. (2015).
Global optimality in tensor factorization, deep learning, and beyond.
CoRR, abs/1506.07540.

[Koh and Liang, 2017] Koh, P. W. and Liang, P. (2017).
Understanding black-box predictions via influence functions.
In Precup, D. and Teh, Y. W., editors, *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1885–1894, International Convention Centre, Sydney, Australia. PMLR.

[Laine and Aila, 2017] Laine, S. and Aila, T. (2017).
Temporal ensembling for semi-supervised learning.
In *International Conference on Learning Representations (ICLR)*.

[Mallat, 2012] Mallat, S. (2012).
Group invariant scattering.
Communications in Pure and Applied Mathematics, 10:1331–1398.

[Sajjadi et al., 2016] Sajjadi, M., Javanmardi, M., and Tasdizen, T. (2016).
Regularization with stochastic transformations and perturbations for deep semi-supervised learning.
In *Advances in Neural Information Processing Systems (NIPS)*.

[Tarvainen and Valpola, 2017] Tarvainen, A. and Valpola, H. (2017).
Mean teachers are better role models: Weight-averaged consistency targets improve semi-supervised deep learning results.
In *Advances in Neural Information Processing Systems (NIPS)*.

[Zhang et al., 2017] Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2017).
Understanding deep learning requires rethinking generalization.