Recent Generalization Bound for DNNs

Shirin Goshtasbpour

May 28th

Outline

- Why leanring with DNNs is surprizing?
- Basic concepts (lots of maths 7-10)
- Some of the recent bounds on generalization of DNNs
- Deriving bound using uniform convergence is hard (Nagarajan '19)

Problem Setting

- Dataset $(\mathbf{x}_i, y_i)_{i=1}^n \sim \mathcal{D}^n$
- Function class $\mathcal{F}_\Theta = \{f_\theta : \mathcal{X} \to \mathcal{Y} | \theta \in \Theta\}$
- Loss function $\mathcal{L}(\hat{y}, y)$

Problem Setting

- Dataset $(\mathbf{x}_i, y_i)_{i=1}^n \sim \mathcal{D}^n$
- Function class $\mathcal{F}_\Theta = \{f_\theta : \mathcal{X} \to \mathcal{Y} | \theta \in \Theta\}$
- Loss function $\mathcal{L}(\hat{y}, y)$
- Optimal Risk

$$f_{\theta^*} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(\mathbf{x}, y) \sim \mathcal{D}}[\mathcal{L}(f_\theta(\mathbf{x}), y)]$$

Problem Setting

- Dataset $(\mathbf{x}_i, y_i)_{i=1}^n \sim \mathcal{D}^n$
- Function class $\mathcal{F}_\Theta = \{f_\theta : \mathcal{X} \to \mathcal{Y} | \theta \in \Theta\}$
- Loss function $\mathcal{L}(\hat{y}, y)$
- Optimal Risk

$$f_{\theta^*} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(\mathbf{x}), y)]$$

- Empirical Risk Minimization (ERM)

$$f_{\hat{\theta}} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta(\mathbf{x}_i), y_i)$$

Problem Setting

- Dataset $(\mathbf{x}_i, y_i)_{i=1}^n \sim \mathcal{D}^n$
- Function class $\mathcal{F}_\Theta = \{f_\theta : \mathcal{X} \to \mathcal{Y} | \theta \in \Theta\}$
- Loss function $\mathcal{L}(\hat{y}, y)$
- Optimal Risk

$$f_{\theta^*} := \arg\min_{\theta \in \Theta} \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_\theta(\mathbf{x}), y)]$$

- Empirical Risk Minimization (ERM)

$$f_{\hat{\theta}} := \arg\min_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta(\mathbf{x}_i), y_i)$$

- **Excess Risk**

$$\mathcal{E}(f_{\hat{\theta}}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat{\theta}}(\mathbf{x}), y) - \mathcal{L}(f_{\theta^*}(\mathbf{x}), y)]$$

- Generalization $\mathcal{E}(f_{\hat{\theta}}) \xrightarrow{n\to\infty} 0$

Optimization Landscape

- Highly non-convex loss function
- Possibly lots of saddle points and local optimas

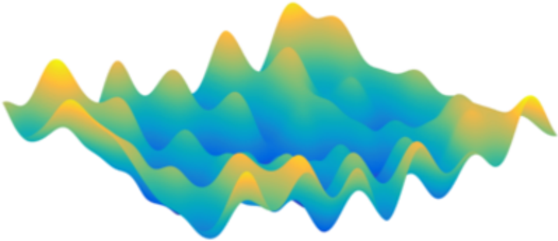


Figure 1: Loss Landscape

Optimization Landscape

- Highly non-convex loss function
- Possibly lots of saddle points and local optimas

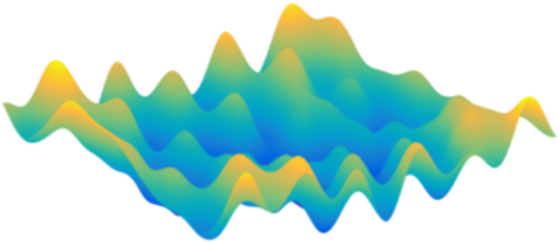
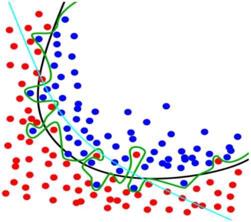


Figure 1: Loss Landscape

- Yet, SGD finds a solution with **low empirical risk**

Underfitting and Overfitting



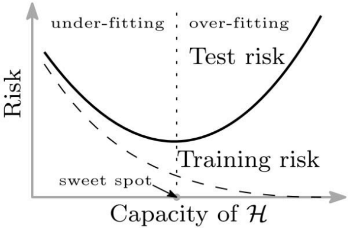
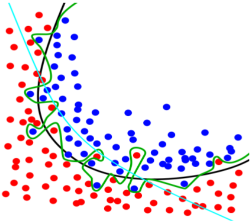


Figure 2: (cyan) Very small and (green) very large number of params (right) overfitting with higher capacity

Underfitting and Overfitting



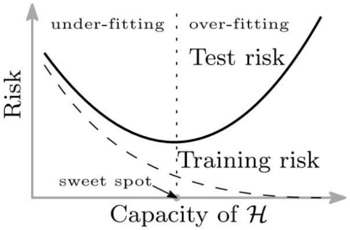


Figure 2: (cyan) Very small and (green) very large number of params (right) overfitting with higher capacity

- Traditionally handled with regularization

$$f_{\hat{\theta}} := \underset{\theta \in \Theta}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(f_\theta(\mathbf{x}_i), y_i) + \lambda \|\theta\|$$

Observations Contradict Traditional Beliefs in ML

- DNNs' capacity is enough to memorize random data (Zhang '18)

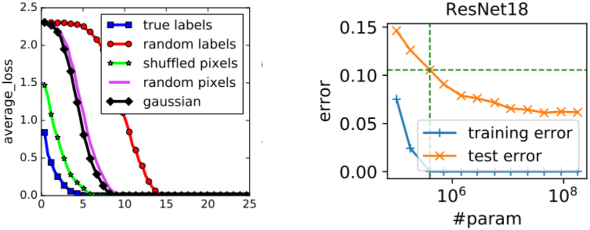
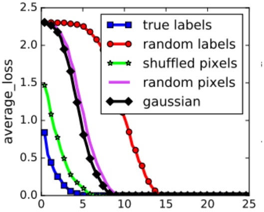


Figure 3: (left) Training classification loss on CIFAR10 (right) better generalization with more params

Observations Contradict Traditional Beliefs in ML

- DNNs' capacity is enough to memorize random data (Zhang '18)



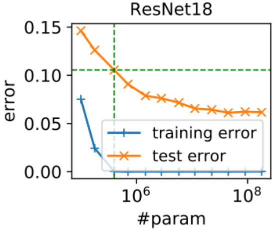


Figure 3: (left) Training classification loss on CIFAR10 (right) better generalization with more params

- Yet, SGD finds a solution that Generalizes to unseen data

Training NN with SGD on **real data** induces regularization

Generalization Gap and Uniform Bound

$$\mathcal{E}(f_{\hat{\theta}}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat{\theta}}(\mathbf{x}), y) - \mathcal{L}(f_{\theta^*}(\mathbf{x}), y)]$$
$$\leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat{\theta}}(\mathbf{x}), y)] - \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f_{\hat{\theta}}(\mathbf{x}_i), y_i) + \epsilon$$

Generalization Gap and Uniform Bound

$$\mathcal{E}(f_{\hat{\theta}}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat{\theta}}(\mathbf{x}), y) - \mathcal{L}(f_{\theta^*}(\mathbf{x}), y)]$$
$$\leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat{\theta}}(\mathbf{x}), y)] - \frac{1}{n}\sum_{i=1}^{n} \mathcal{L}(f_{\hat{\theta}}(\mathbf{x}_i), y_i) + \epsilon$$

When $n \to \infty$

- ϵ converges with central limit theorem
- Excess risk converges with the same rate as Generalization Gap
- Not i.i.d samples

Generalization Gap and Uniform Bound

$$\mathcal{E}(f_{\hat\theta}) := \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat\theta}(\mathbf{x}),y) - \mathcal{L}(f_{\theta^*}(\mathbf{x}),y)]$$
$$\leq \mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\hat\theta}(\mathbf{x}),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f_{\hat\theta}(\mathbf{x}_i),y_i) + \epsilon$$

When $n \to \infty$

- ϵ converges with central limit theorem
- Excess risk converges with the same rate as Generalization Gap
- Not i.i.d samples

Uniform bound

Orange term $\leq \sup_{\theta\in\Theta}\mathbb{E}_{(\mathbf{x},y)\sim\mathcal{D}}[\mathcal{L}(f_{\theta}(\mathbf{x}),y)] - \frac{1}{n}\sum_{i=1}^{n}\mathcal{L}(f_{\theta}(\mathbf{x}_i),y_i)$

Rademacher Complexity and Uniform Convergence

- Rademacher complexity reflects richness of function space
- $\epsilon_i \in \{-1, 1\}$ with probability $\frac{1}{2}$

$$\mathcal{R}(\mathcal{F}_\Theta) := \mathbb{E}_{\mathbf{x}_i, y_i, \epsilon_i}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_\theta(\mathbf{x}_i)\right]$$

Rademacher Complexity and Uniform Convergence

- Rademacher complexity reflects richness of function space
- $\epsilon_i \in \{-1, 1\}$ with probability $\frac{1}{2}$

$$\mathcal{R}(\mathcal{F}_\Theta) := \mathbb{E}_{\mathbf{x}_i, y_i, \epsilon_i}\left[\sup_{\theta \in \Theta} \frac{1}{n} \sum_{i=1}^{n} \epsilon_i f_\theta(\mathbf{x}_i)\right]$$

Uniform Convergence Theorem
For b-uniformly bounded $\mathcal{L}$, with probability $> 1 - \delta$

$$\text{Uniform bound } \leq 2\mathcal{R}(\mathcal{L} \circ \mathcal{F}_\Theta) + \sqrt{\frac{2\log(1/\delta)}{n}}$$

- If $\mathcal{R}(\mathcal{L} \circ \mathcal{F}_\Theta) = o(1)$ then uniform bound $\xrightarrow{a.s.}$ 0 exponentially
- Rademacher complexity is tight

VC Dimension

- Largest n s.t. there is a $(\mathbf{x}_i)_{i=1}^n \in \mathcal{X}$ that we can assign any binary label $\{0,1\}^n$ to it using functions in $\mathcal{F}_\Theta$

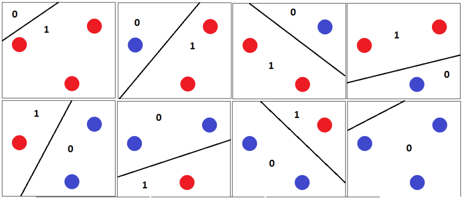


Figure 4: 2d linear classifier can shatter 3 points

VC Dimension

- Largest n s.t. there is a $(\mathbf{x}_i)_{i=1}^n \in \mathcal{X}$ that we can assign any binary label $\{0,1\}^n$ to it using functions in $\mathcal{F}_\Theta$

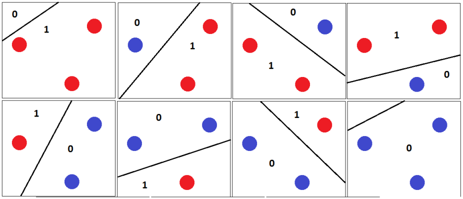


Figure 4: 2d linear classifier can shatter 3 points

> **Uniform VC Bound**
> For binary $\mathcal{L}$ (e.g. $yf(\mathbf{x}) < 0$) $\qquad \mathcal{R}(\mathcal{L} \circ \mathcal{F}_\Theta) \leq 2\sqrt{\frac{d_{\mathrm{VC}} \log(n+1)}{n}}$

- Very loose for rich function classes

Harvey's asymtotically tight VC Dimension Bound

Fully connected DNN

- Total number of parameters (weights and biases) M
- Depth L
- Number of neurons U
- $(p+1)$ piece polynomial activations with degree less than d

$$c.ML\log(M/L) \leq d_{\mathrm{VC}}(M,L) \leq C.ML\log M$$

- Tight for ReLU and leaky ReLU activations

$$d_{\mathrm{VC}}(M,L) = \mathcal{O}(MU\log(d+1)p)$$

Margin Bound

Classification margin

$$f(\mathbf{x}_i)_{y_i} - \max_{j \neq y_i} f(\mathbf{x}_i)_j$$

- Margin loss

$$\mathcal{L}_\gamma(f_{\hat{\theta}}, (\mathbf{x}_i, y_i)_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}[f(\mathbf{x}_i)_{y_i} - \max_{j \neq y_i} f(\mathbf{x}_i)_j \leq \gamma]$$

Margin Bound

Classification margin

$$f(\mathbf{x}_i)_{y_i} - \max_{j \neq y_i} f(\mathbf{x}_i)_j$$

- Margin loss

$$\mathcal{L}_\gamma(f_{\hat{\theta}}, (\mathbf{x}_i, y_i)_{i=1}^n) := \frac{1}{n} \sum_{i=1}^n \mathbb{1}[f(\mathbf{x}_i)_{y_i} - \max_{j \neq y_i} f(\mathbf{x}_i)_j \leq \gamma]$$

Margin Bound
With probability $> 1 - \delta$

$$\mathbb{P}[\arg\max_j f(\mathbf{x})_j \neq y] \leq \mathcal{L}_\gamma(f_{\hat{\theta}}, \ldots) + \frac{2}{\gamma} \mathcal{R}(\mathcal{F}_\Theta) + \mathcal{O}\left(\sqrt{\frac{\log(1/\delta)}{n}}\right)$$

Explaining SGD's Bias

After training DNNs on **real data**

- Aggregated updates have small singular values
- DNNs have bounded lipschitz constant
- Weights don't change much from initialization
    - Despite the long training time :)

$\implies$ Investigate complexity when wieghts have bounded norms

Feed Forward Network with Constraints (Neyshabur '18)

- $\|\mathbf{x}_i\|_2 \leq B$
- Depth L and width h
- Weight $\mathbf{W}_i$ for ith layer
- ReLU activations

Rademacher complexity of DNN with constrained lipschitz constant

$$\mathcal{O}\left(\frac{BL\sqrt{h}}{\sqrt{n}}\left(\prod_{i=1}^{L}\|\mathbf{W}_i\|_\sigma\right)\left(\sum_{i=1}^{L}\frac{\|\mathbf{W}_i-\mathbf{W}_{i,0}\|_F^2}{\|\mathbf{W}_i\|_\sigma^2}\right)^{1/2}\right)$$

Feed Forward Network with Constraints (Bartlett '17)

- $\|\mathbf{x}_i\|_2 \leq B$
- Depth L and width h
- Weight $\mathbf{W}_i$ for ith layer
- ρ_i-Lipschitz activations at ith layer

Rademacher complexity of DNN with constrained lipschitz constant

$$\mathcal{O}\left(\frac{B}{\sqrt{n}}\left(\prod_{i=1}^{L}\rho_i\|\mathbf{W}_i\|_\sigma\right)\left(\sum_{i=1}^{L}\frac{\|\mathbf{W}_i-\mathbf{W}_{i,0}\|_{2,1}^{2/3}}{\|\mathbf{W}_i\|_\sigma^{2/3}}\right)^{3/2}\right)$$

Feed Forward Network with Constraints (Neyshabur '19)

- $\|\mathbf{x}_i\|_2 \leq B$
- Two-layer ReLU NN
- First and second layer weights **U** and **V**
- k classes

Rademacher complexity of DNN with constrained distance norms

$$\mathcal{O}\left(\frac{B\sqrt{k}}{\sqrt{n}}\|\mathbf{V}\|_F\left(\|\mathbf{U}-\mathbf{U}_0\|_F+\|\mathbf{U}_0\|_\sigma\right)+\sqrt{\frac{h}{n}}\right)$$

Nagarajan's Arguments

- Consider high probability datasets $\mathcal{S}_\delta$
- Classifiers $h \in \mathcal{H}_\delta$ trained on $\mathcal{S}_\delta$ have low generalization error $\leq \epsilon$ w.h.p

If for every $h \in \mathcal{H}_\delta$ there is a dataset $S^{-}(h) \in \mathcal{S}_\delta$ that is misclassified w.h.p. $\implies$ $|\text{uniform bound}| \geq 1 - \epsilon$

Nagarajan's Arguments

- Consider high probability datasets $\mathcal{S}_\delta$
- Classifiers $h \in \mathcal{H}_\delta$ trained on $\mathcal{S}_\delta$ have low generalization error $\leq \epsilon$ w.h.p

> If for every $h \in \mathcal{H}_\delta$ there is a dataset $S^-(h) \in \mathcal{S}_\delta$ that is misclassified w.h.p. $\implies |\text{uniform bound}| \geq 1 - \epsilon$

- Uniform convergence is all we have
  - Rademacher complexity
  - VC dim
  - PAC learning
  - Covering number

Connection to Adversarial Samples

- S^{-} is adversarial dataset for h
- In high dimensions almost all training samples can fool the classifier

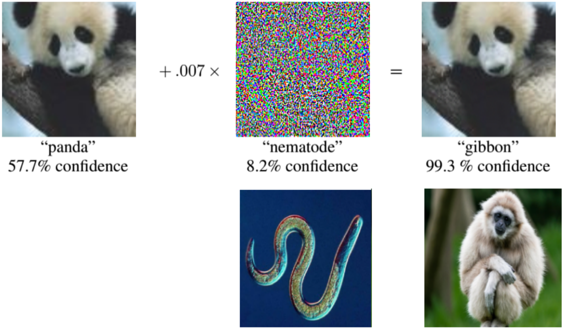


Figure 5: Just mutate with the right gene

- Is it high probability or does it fall off the data's manifold?

Nagarajan's Experiments

- Train a deep fc network with SGD batch size 1

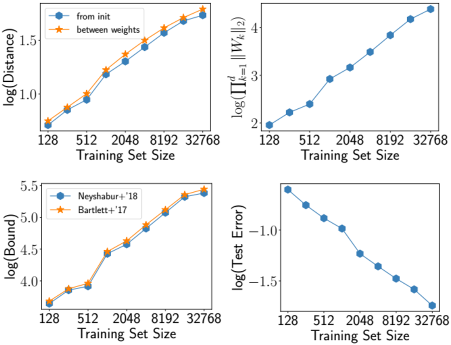


Figure 6: More noise stronger adversaries

DNNs have nothing better to do than memorize the noise in data

Compressability of SGD solution on real data

MNIST Classifier

- Removing unimportant singular values from weight updates

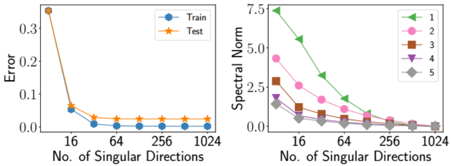


Figure 7: Too many useless singular directions

Compressed networks are more robust to adversarial examples

:)

Thank You

References

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. *Understanding deep learning requires rethinking generalization.* In Proceedings of the International Conference on Learning Representations (ICLR), 2017

Harvey, N., Liaw, C., and Mehrabian, A. "Nearly-tight vcdimension bounds for piecewise linear neural networks". In Proceedings of the 30th Conference on Learning Theory, COLT 2017, 2017

Neyshabur, B., Bhojanapalli, S., McAllester, D., and Srebro, N. "A pac-bayesian approach to spectrally-normalized margin bounds for neural networks". International Conference on Learning Representations, 2018

Bartlett, P. L., Foster, D. J., and Telgarsky, M. J. "Spectrallynormalized margin bounds for neural networks". In Advances in Neural Information Processing Systems 2017

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. "The role of over-parametrization in generalization of neural networks". In International Conference on Learning Representations (ICLR), 2019

Nagarajan, Vaishnavh, and J. Zico Kolter. *Uniform convergence may be unable to explain generalization in deep learning.* NeurIPS 2019.