Seminar on Deep Learning Theory **IST Austria**

Generalisation in Deep Learning

Presenter: Mary Phuong *26 April 2019*

1 Why theory of deep learning?

1.1 Setup

Task:

- input-output pairs $(\mathbf{x}, y) \sim D$ on $\mathcal{X} \times \{+1, -1\}$,
- goal: find hypothesis $h : \mathcal{X} \to \{+1, -1\}$,
- quality measured by risk $R(h) := \mathbb{E}_{(\mathbf{x},y)\sim D}[\ell(h(\mathbf{x}), y)]$.

How do we find the hypothesis?

- access to training data $S = \{(\mathbf{x}_1, y_1), \ldots, (\mathbf{x}_n, y_n)\}$,
- learning algorithm $A : S \mapsto h$,
- empirical risk $\hat{R}_S(h) := \frac{1}{n} \sum_{(\mathbf{x},y)\in S} \ell(h(\mathbf{x}), y)$.

Goal of learning theory: Understand generalisation, as a function of the learning algorithm A.
Measure of generalisation: generalisation gap, $\Delta(h) = R(h) - \hat{R}_S(h)$.
Why?

- To have good, easily communicable mental models of learning algorithms; To drive algorithm design.
- Not our primary goal: to obtain upper bounds on the generalisation gap (we can keep a held-out set). Bounds symbolise and make precise conceptual models, but are not literally of interest by themselves.

1.2 Classical learning theory

- Finite hypothesis set. If $A(S) \in \mathcal{H}$, then with prob. $1 - \delta$,
$$\Delta(h) \le \sqrt{\frac{\log|\mathcal{H}| + \log(1/\delta)}{n}}. \quad (1)$$
Example: low-precision neural net [1].
- VC dimension (number of examples needed to force training error > 0).
Let $\ell(\hat{y}, y) = \mathbb{1}\{\hat{y} = y\}$. If $A(S) \in \mathcal{H}$, then with prob. $1 - \delta$,
$$\Delta(h) \le C\sqrt{\frac{\mathrm{VCdim}(\mathcal{H}) + \log(1/\delta)}{n}}. \quad (2)$$
For fully-connected ReLU networks with p parameters and L layers, $\mathrm{VCdim}(\mathcal{H}) = \tilde{O}(pL)$ [5].
- Rademacher complexity. If $A(S) \in \mathcal{H}$, then with prob. $1 - \delta$,
$$\Delta(h) \le 2\mathcal{R}_{D^n}(\ell \circ \mathcal{H}) + \sqrt{\frac{\log(1/\delta)}{n}}. \quad (3)$$

1.3 Deep learning contradicts learning theory

These bounds don't apply to deep learning.

- Symptom: bounds are extremely loose.
- Underlying problem: the bounds are conceptually wrong.
- According to classical learning theory: **good generalisation** $\approx$ **small capacity / complexity**.
  Deep learning: good generalisation **and** massive capacity.
- More generally, classical theory frames generalisation as a function of the model class: The right-hand side of classical bounds depends only on $\mathcal{H}$. It may be reasonable to say h generalises because it's a linear model. Is it reasonable to say h generalises because it's a neural network?
  $\Longrightarrow$ Interaction with optimisation!

Q: What would be a good mental model of generalisation in deep learning?

2 Proposed solutions

2.1 Stability

Intuition: SGD, run for a short time, isn't very sensitive to single examples in the training set, therefore it cannot overfit. The complexity of the hypothesis class doesn't matter when trained this way.
Denote $S^{\setminus i} := S \setminus \{(\mathbf{x}_i, y_i)\}$.

Definition 1 (Uniform stability)**.** *A learning algorithm A is β-*uniformly stable *if, for all training sets S and $i \in [n]$, the hypotheses $h := A(S)$ and $h' := A(S^{\setminus i})$ satisfy*

$$\sup_{\mathbf{x},y} |\ell(h(\mathbf{x}), y) - \ell(h'(\mathbf{x}), y)| \leq \beta. \tag{4}$$

Lemma 1 (Bousquet et al. [3])**.** *Let A be β-uniformly stable and let $|\ell| \leq M$. Then with prob. $1-\delta$,*

$$\Delta(h) \leq 2\beta + (4n\beta + M)\sqrt{\frac{\log(1/\delta)}{2n}}. \tag{5}$$

Consider running SGD on $f(\boldsymbol{\theta}) = \frac{1}{n}\sum_{i=1}^{n} f(\boldsymbol{\theta} \,|\, \mathbf{x}_i, y_i)$.
(A function $f : \boldsymbol{\Theta} \to \mathbb{R}$ is γ-smooth if for all $\mathbf{u}, \mathbf{v} \in \boldsymbol{\Theta}$ we have $\|\nabla f(\mathbf{u}) - \nabla f(\mathbf{v})\| \leq \gamma\|\mathbf{u} - \mathbf{v}\|$.
A function $f : \boldsymbol{\Theta} \to \mathbb{R}$ is L-Lipschitz if for all $\mathbf{u}, \mathbf{v} \in \boldsymbol{\Theta}$ we have $|f(\mathbf{u}) - f(\mathbf{v})| \leq L\|\mathbf{u} - \mathbf{v}\|$.)

Theorem 1 (Hardt et al. [4])**.** *Let $f(\,.\,|\,\mathbf{x}, y)$ be γ-smooth and L-Lipschitz. Then SGD run with step sizes $\alpha_t \leq c/t$, is β-uniformly stable with*

$$\beta \leq \frac{C(\gamma, c, L)}{n} \cdot T^{\gamma c/(\gamma c+1)} \tag{6}$$

Remarks:

- It is intuitive that stability is sufficient for generalisation, but it may be too strong.
- Explains DL? Anything trained by SGD generalises well.

2.2 Margins and norms

Intuition: not all neural nets generalise well, but those that have the following properties do.

- High confidence in (correct) predictions.
- Small weights (small complexity).

(SGD preferentially finds these.)
Define

- a neural network $f_{\mathbf{W}}(\mathbf{x}) := \sigma(\mathbf{W}_L\sigma(\mathbf{W}_{L-1}\cdots\sigma(\mathbf{W}_1\mathbf{x})\cdots))$, where $\mathbf{W} = (\mathbf{W}_1, \ldots, \mathbf{W}_L)$ are the weights and σ is the component-wise ReLU,
- a network's width $d :=$ max. number of neurons in a single layer,
- scaling of the data $\|\mathbf{x}\|^2 := \frac{1}{n}\sum_i \|\mathbf{x}_i\|^2$,
- (small weights) *spectral complexity* of the network $f_{\mathbf{W}}$,

$$C_{\mathbf{W}} \approx L^{3/2}d^{1/2}\left(\prod_{l=1}^{L}\|\mathbf{W}_l\|_2\right), \tag{7}$$

- (high confidence) *margin* $\gamma(\mathbf{x}, y) := f_{\mathbf{W}}(\mathbf{x})[y] - \max_{y'\neq y} f_{\mathbf{W}}(\mathbf{x})[y']$.

Theorem 2 (Bartlett et al. [2]). *With prob. $1-\delta$, the following holds for all $\gamma > 0$,*

$$R_{01}(f_{\mathbf{W}}) - \hat{R}_\gamma(f_{\mathbf{W}}) \leq \tilde{O}\left(\frac{\|\mathbf{x}\|\cdot C_{\mathbf{W}}}{\gamma\sqrt{n}}\log(d) + \sqrt{\frac{\log(1/\delta)}{n}}\right), \tag{8}$$

where $\hat{R}_\gamma(f) := \frac{1}{n}\sum_{i=1}^{n}\mathbb{1}\{\gamma(\mathbf{x}, y) \leq \gamma\}$.

2.3 Flat minima

Intuition: hypothesis that generalises well should be resilient to small weight perturbations.
SGD is more likely to find a flat minimum.
No established definition of flatness / sharpness, one possible is this.
Define

- weight perturbation $\nu \sim \mathcal{N}(\mathbf{0}, \tau\mathbf{I})$,
- *expected sharpness* $S_{\mathbf{W}} := \mathbb{E}_\nu\left[\hat{R}(f_{\mathbf{W}+\nu})\right] - \hat{R}(f_{\mathbf{W}})$.

Theorem 3 (Neyshabur et al. [6]). *With prob. $1-\delta$,*

$$\mathop{\mathbb{E}}_{\nu}[R(f_{\mathbf{W}+\nu})] - \hat{R}(f_{\mathbf{W}}) \leq S_{\mathbf{W}} + 4\sqrt{\frac{\|\mathbf{W}\|_2^2/2\tau^2 + \log(2n) + \log(1/\delta)}{n}}. \tag{9}$$

Notes:

- Caveat: Bound on perturbed risk!
- Additional norm control.
- Relationship between norm and sharpness via τ.

References

[1] Sanjeev Arora, Rong Ge, Behnam Neyshabur, and Yi Zhang. Stronger generalization bounds for deep nets via a compression approach. *arXiv:1802.05296*, 2018.

[2] Peter L Bartlett, Dylan J Foster, and Matus J Telgarsky. Spectrally-normalized margin bounds for neural networks. *NIPS*, 2017.

[3] Olivier Bousquet and André Elisseeff. Stability and generalization. *JMLR*, 2(Mar):499–526, 2002.

[4] Moritz Hardt, Benjamin Recht, and Yoram Singer. Train faster, generalize better: Stability of stochastic gradient descent. *ICML*, 2016.

[5] Nick Harvey, Chris Liaw, and Abbas Mehrabian. Nearly-tight VC-dimension bounds for piecewise linear neural networks. *COLT*, 2017.

[6] Behnam Neyshabur, Srinadh Bhojanapalli, David McAllester, and Nati Srebro. Exploring generalization in deep learning. *NIPS*, 2017.