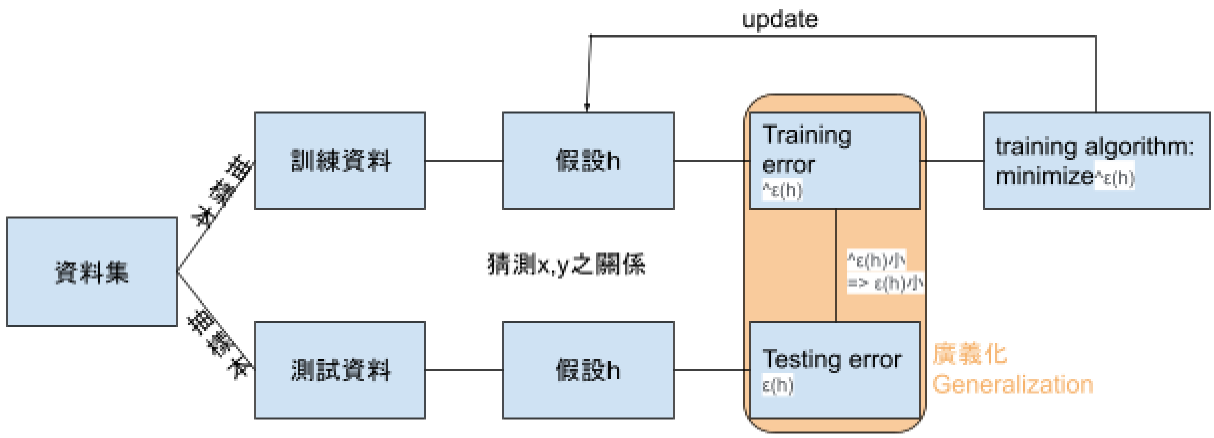
人工智慧應用專題-個人報告二

Mark Chang - Deep Learning Theory　　　　　Author: Eason Chu

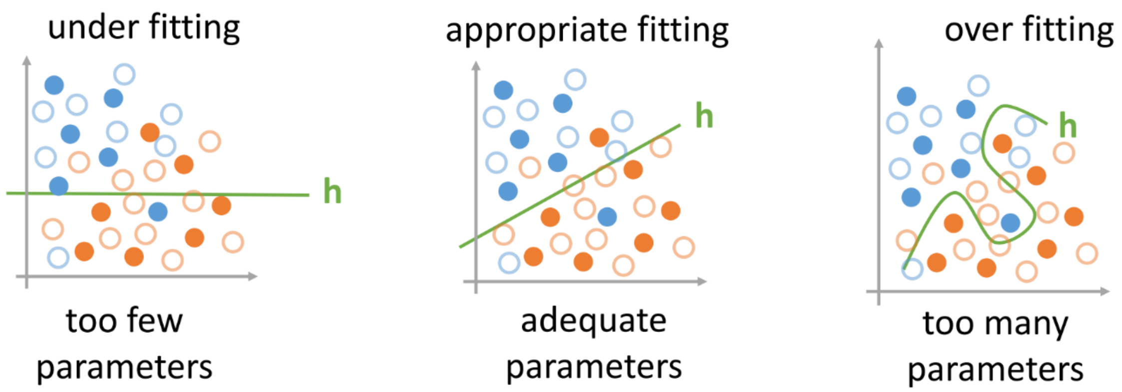
備註：此份報告為課堂筆記，內容多擷取Mark Chang老師上課PPT。

Part I. Learning Theory



OverFitting(過適): ^ε(h) 小 但ε(h)很大

過多參數→High VC Dimension→OverFitting



(來源: Mark Chang老師上課PPT)

VC Bound不等式:

$$\epsilon(h) \leq \hat{\epsilon}(h) + \sqrt{\frac{8}{n}\log\left(\frac{4(2n)^d}{\delta}\right)}$$

numbef of
training instances

VC Dimension
(model complexity)

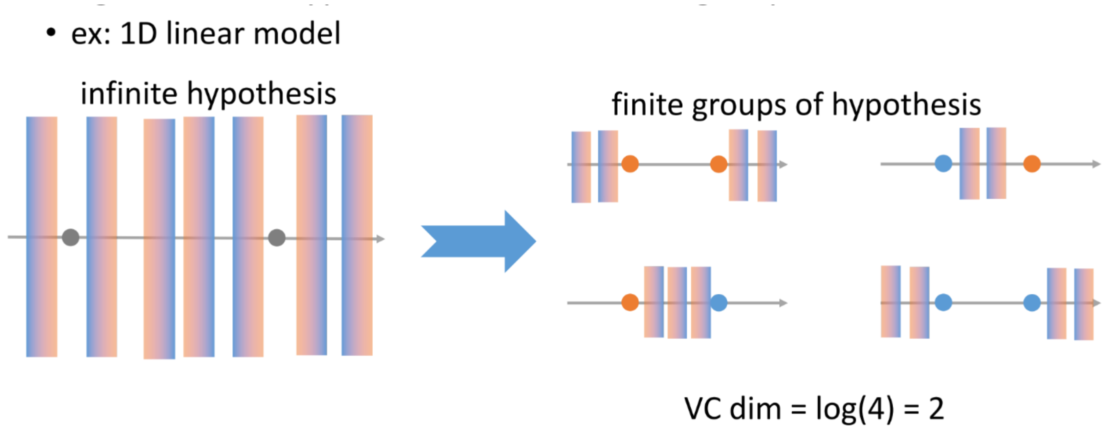
(來源: Mark Chang老師上課PPT)

若Training error跟Testing data越多→Testing error越小→不會OverFitting

VC Dimension:

1.關乎模型複雜度

老師上課所舉之例子:

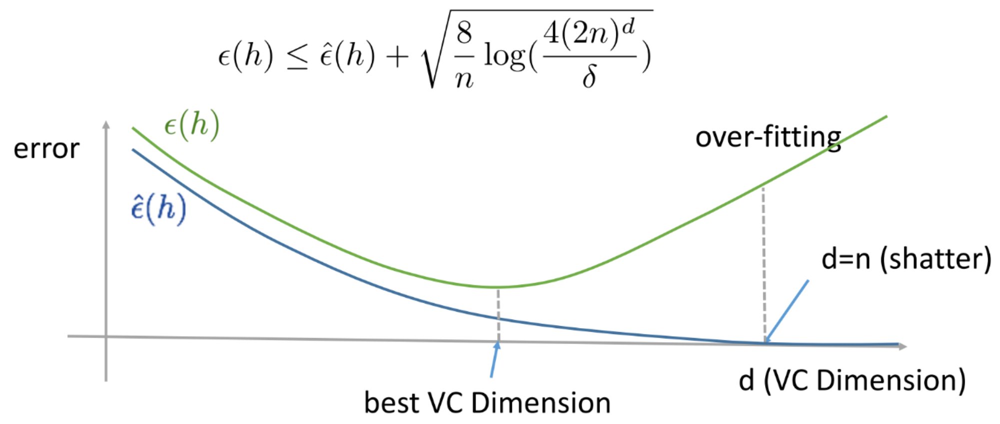


2.d(H) Formula

$$\mathrm{d}(H) = \max\{n : \tau_H(n) = 2^n\}$$

尋找最佳VC Dimension:

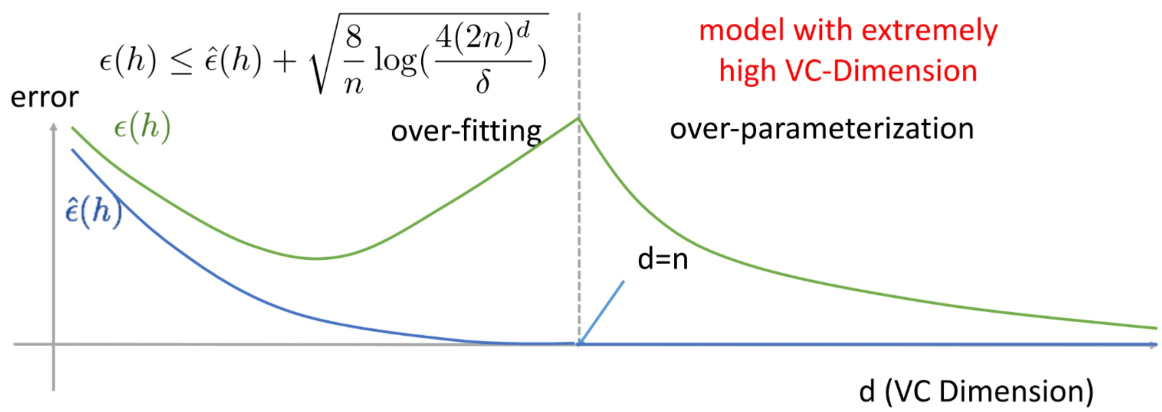
- For a given dataset (n is constant), search for the best VC Dimension



Model VC dimension過高 → OverFitting
→降低VC dimension → 找出合適Model

Part II. Deep Learning

Generalization in Deep Learning:



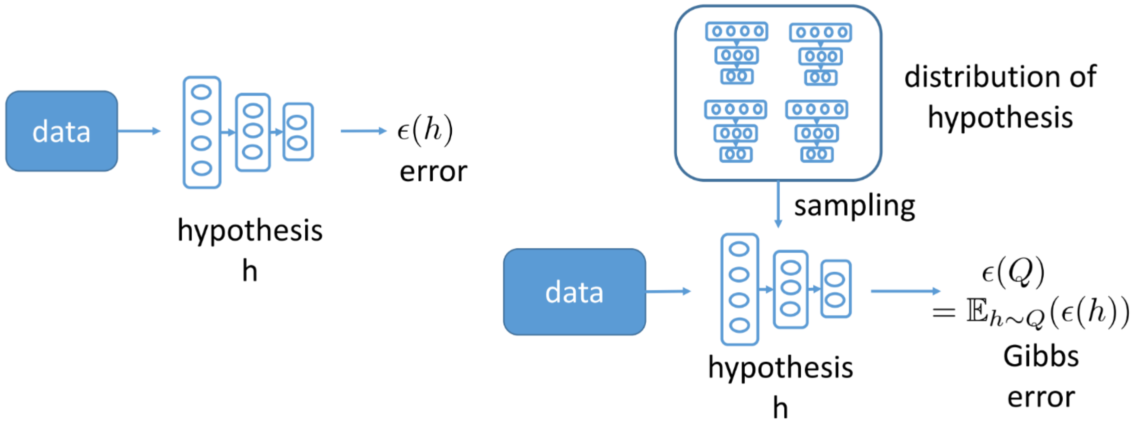
發現VC Dimension 非常高時 → OverFitting到達極值後會逐漸降低→得到較好的VC Dimension
結論: Deep Learning中，參數越多效果越好

小結 : 深度學習中，OverFitting與模型關聯較小，但與資料的組成關聯大
→若無法訓練出適當的應先從資料蒐集下手

Part III. PAC-Bayesian Learning

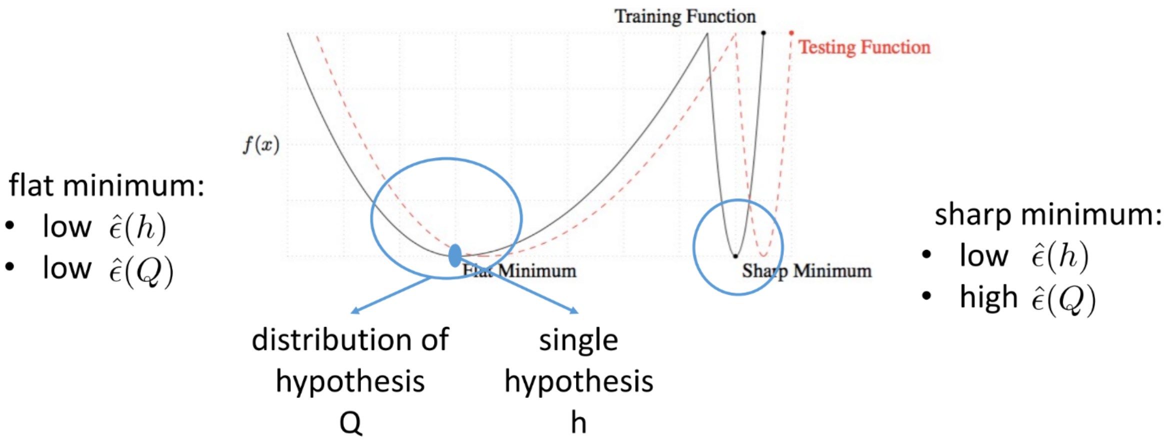
2 Major Theory For Deep Learining:

- Deterministic Model
- Stochastic Model (Gibbs Classifier)



PAC-Bayesian Bound:

1.Graph of the sharpness of local minimums



若把周圍資料也考慮進去則較準確

2.PAC-Bayesian Bound不等式:

$$\epsilon(Q) \leq \hat{\epsilon}(Q) + \sqrt{\frac{KL(Q\|P) + \log(\frac{n}{\delta}) + 2}{2n - 1}}$$

KL divergence between P and Q

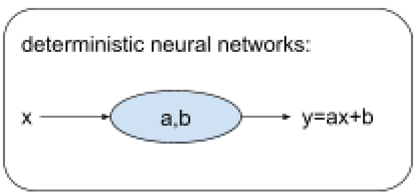
number of training instances

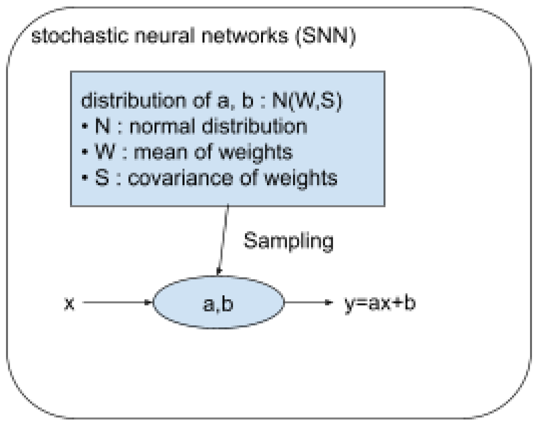
由此可知，PAC-Bayesian Bound較依賴資料(Data-dependent)

- High VC Dimension, but clean data
-> low KL (Q||P)

- High VC Dimension, but noisy data
-> high KL(Q||P)

PAC-Bayesian Bound for Deep Learning(Stochastic):





$$KL(\hat{\epsilon}(Q)\|\epsilon(Q)) \leq \frac{KL(Q\|P) + \log(\frac{n}{\delta})}{n - 1}$$

Q: SNN after training

P: SNN before training

PAC-Bayesian Bound for Deep Learning(Deterministic):

- Constrain on sharpness: given a margin γ > 0, if Q satisfy:

$$P_{h'\sim Q}\left\{\sup_{x\in X}\|h'(x)-h(x)\|_\infty \le \frac{\gamma}{4}\right\} \ge \frac{1}{2}$$

- With 1-δ probability, the following inequality is satisfied.

$$L_0(h) \le \hat{L_\gamma}(h) + 4\sqrt{\frac{KL(Q\|P)+\log(\frac{6n}{\delta})}{n-1}}$$

- margin loss $L_\gamma(h)$:

$$L_\gamma(h) = P_{(x,y)\sim D}\left\{h(x)[y] \le \gamma + \max_{j\neq y} h(x)[j]\right\}$$

Part IV. How to overcome Overfitting?

$$\epsilon(h) \le \hat{\epsilon}(h) + \sqrt{\frac{8}{n}\log(\frac{4(2n)^d}{\delta})}$$

Traditional Machine Learning

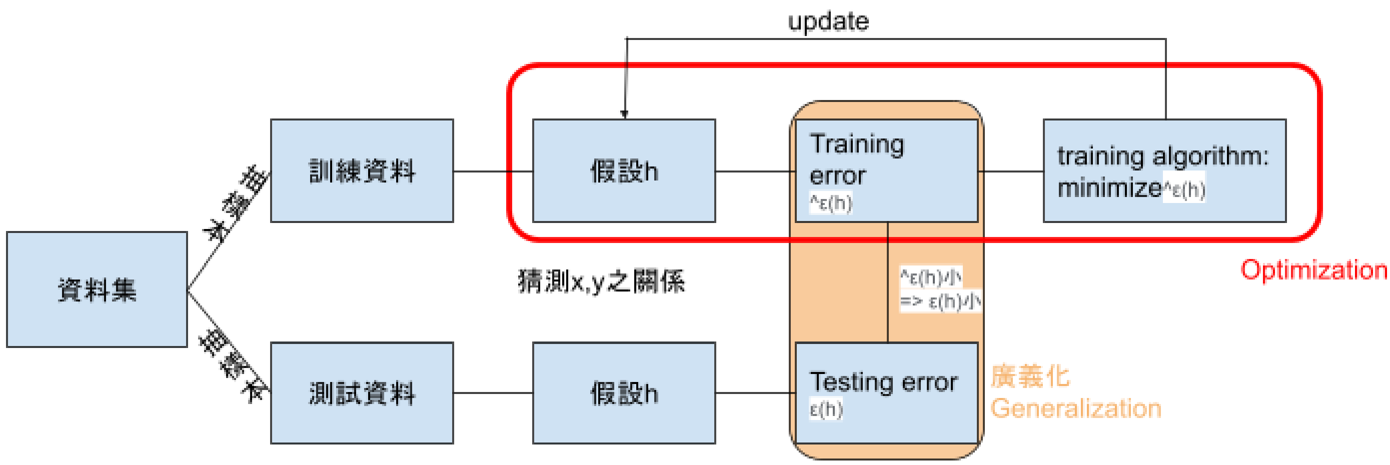
- 減少參數數量
- weight decay
- Early Stop
- 資料增強(data augmentation)

$$KL(\hat{\epsilon}(Q)\|\epsilon(Q)) \le \frac{KL(Q\|P)+\log(\frac{n}{\delta})}{n-1}$$

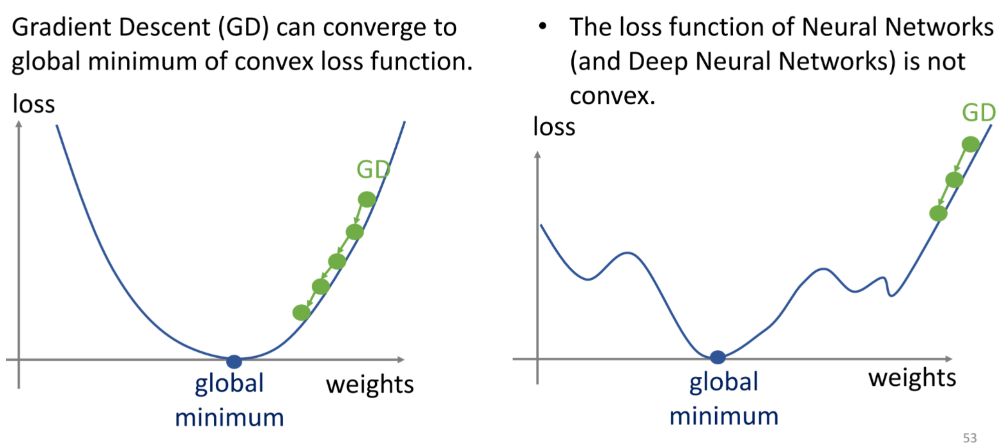
Modern Deep Learning

- Early Stop?
- weight decay?
- 資料增強(data augmentation)?
- 改善資料品質(data quality)
- 從好的資料下手(P: 測試前資料)

Part V. How to overcome Neural Tangent Kernel?



Optimization(最佳化):



Gram Matrix(H∞):

- Gram Matrix H∞ is the Neural Tangent Kernel of two-layer ReLU NN

$$\mathbf{H}_{ij}^{\infty} = \sum_{r=1}^{\infty} \langle \frac{\partial f_{\mathbf{W}}(\mathbf{x}_i)}{\partial \mathbf{w}_r}, \frac{\partial f_{\mathbf{W}}(\mathbf{x}_j)}{\partial \mathbf{w}_r} \rangle$$

Part VI. Analysis of Generalization

Theorem 5.1:

Given a probability δ, suppose that :

1. The dataset $S = \{(\mathbf{x}_i, y_i)\}_{i=1}^n$ samples from $a(\lambda_0, \delta/3, n)$-*non-degenerate* $\mathcal{D}$.
2. $m \geq \kappa^{-2}\mathrm{poly}(n, \lambda_0^{-1}, \delta^{-1})$, where $\kappa = O(\frac{\lambda_0 \delta}{n})$, (over-parameterization)
3. The loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0,1]$ is 1-Lipschitz in the first argument.
4. The two-layer NN $f_{\mathbf{W}(k),\mathbf{a}}$ trained by GD for $k \geq \Omega(\frac{1}{\eta\lambda_0}\log\frac{n}{\delta})$ iterations.

Define $L_{\mathcal{D}}(f_{\mathbf{W}(k),\mathbf{a}}) = \mathbf{E}_{(\mathbf{x},y)\sim\mathcal{D}}\left[\ell\left(f_{\mathbf{W}(k),\mathbf{a}}(\mathbf{x}), y\right)\right]$ as population loss,

with probability at least $1-\delta$:

$$L_{\mathcal{D}}(f_{\mathbf{W}(k),\mathbf{a}}) \leq \sqrt{\frac{2\mathbf{y}^T(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + O\left(\sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n}}\right)$$

Theorem 5.1 v.s. VC Bound v.s. PAC-Bayesian Bound

VC Bound	PAC-Bayesian Bound	Theorem 5.1
$\epsilon(h) \leq \hat{\epsilon}(h) + \sqrt{\frac{8}{n}\log(\frac{4(2n)^d}{\delta})}$	$\epsilon(Q) \leq \hat{\epsilon}(Q) + \sqrt{\frac{KL(Q\|P) + \log(\frac{n}{\delta}) + 2}{2n-1}}$	$L_{\mathcal{D}}(f_{\mathbf{W}(k),\mathbf{a}}) \leq \sqrt{\frac{2\mathbf{y}^T(\mathbf{H}^\infty)^{-1}\mathbf{y}}{n}} + O\left(\sqrt{\frac{\log\frac{n}{\lambda_0\delta}}{n}}\right)$
Only depends on model Can not be applied to over-parameterization NN	Depends on both model and training data Model needs to be trained	Only depends on training data Model doesn't need to be trained Can only be applied to over-parameterized 2-layer ReLU NN

Part VII. Reference

Chang, M. (2022). *NCCU Deep Learning Theory*.