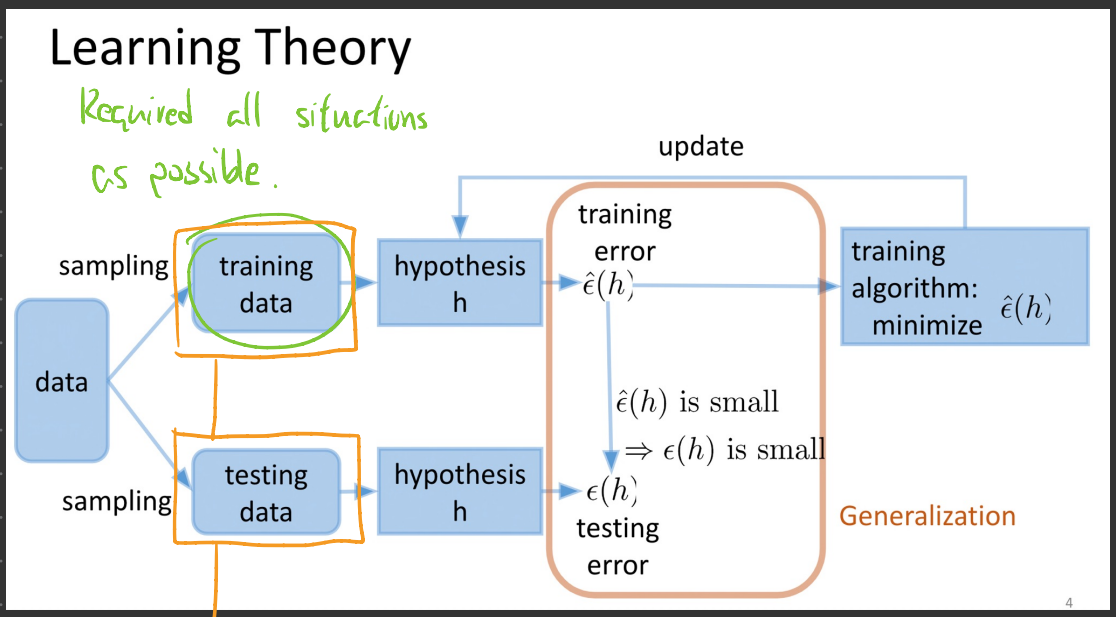
11/18. Fri. 2022

* Learning Theory



They have same distribution

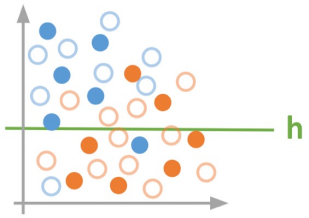
* Overfitting

Learning Theory

- Over fitting : $\hat{\epsilon}(h)$ is small, but $\epsilon(h)$ large
- What cause over-fitting?

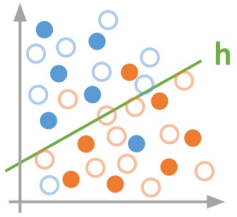
training data
testing data

under fitting



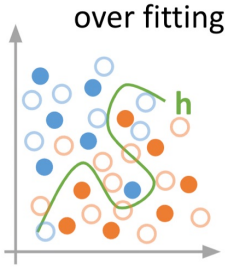
too few parameters

appropriate fitting



adequate parameters

over fitting



too many parameters

← Random Forest 分支過多時

$$\mathcal{E}(h) \le \hat{\mathcal{E}}(h) + \sqrt{\frac{\delta}{n}\log\left(\frac{4(2n)^d}{\delta}\right)}$$

where d is VC Dimension (model. complexity), n : num. of training instances.

n.d 需要平衡

11/18. Fri. 2022

* VC Dimension.

1. $O(W)$, W: num. of params. for linear model
2. $O(LW \cdot \log W)$, L: num. of layer, W: num. of params

full-connected neural network.

- Growth Function

$$\tau(x_1, ..., x_n) = \left| \left\{ (h(x_1), ..., h(x_n)) \middle| h \in H \right\} \right|$$

$$\tau_H(n) = \max_{(x_1,...,x_n)} \tau(x_1, ..., x_n)$$

data samples; hypothesis set

- VC Dimension

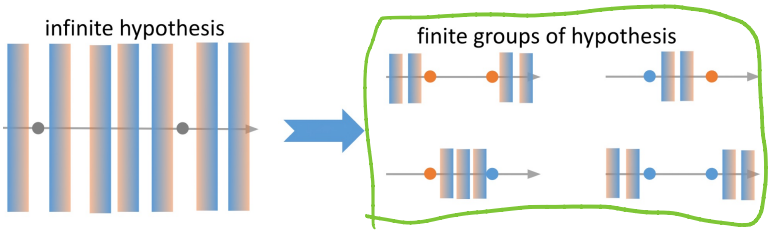
$$d(H) = \max\{n : \tau_H(n) = 2^n\}$$

VC Dimension

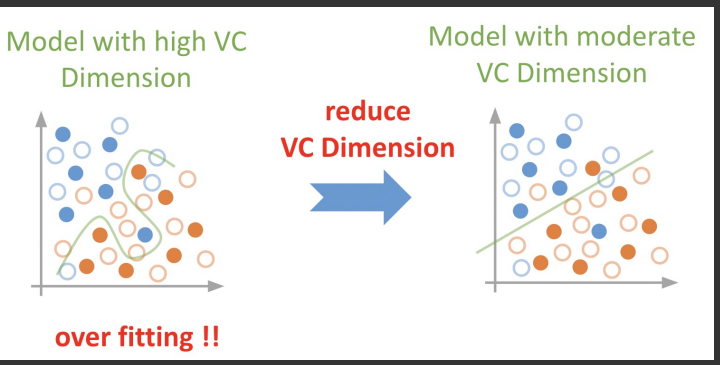
Model Complexity

Categorized the hypothesis set into finite groups

- ex: 1D linear model



* VC Bound.



Ex $\tau(x_1, x_n) = |\{(h(x_1), h(x_n)) | h \in H\}| = 4$

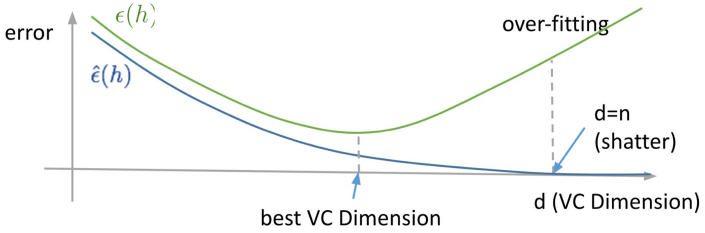
$\{(h(x_1), h(x_n))\} = \{(1,1), (0,1), (1,0), (0,0)\}$

⇒ 可以有幾種狀況的維度

VC Bound

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

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



* Generalization

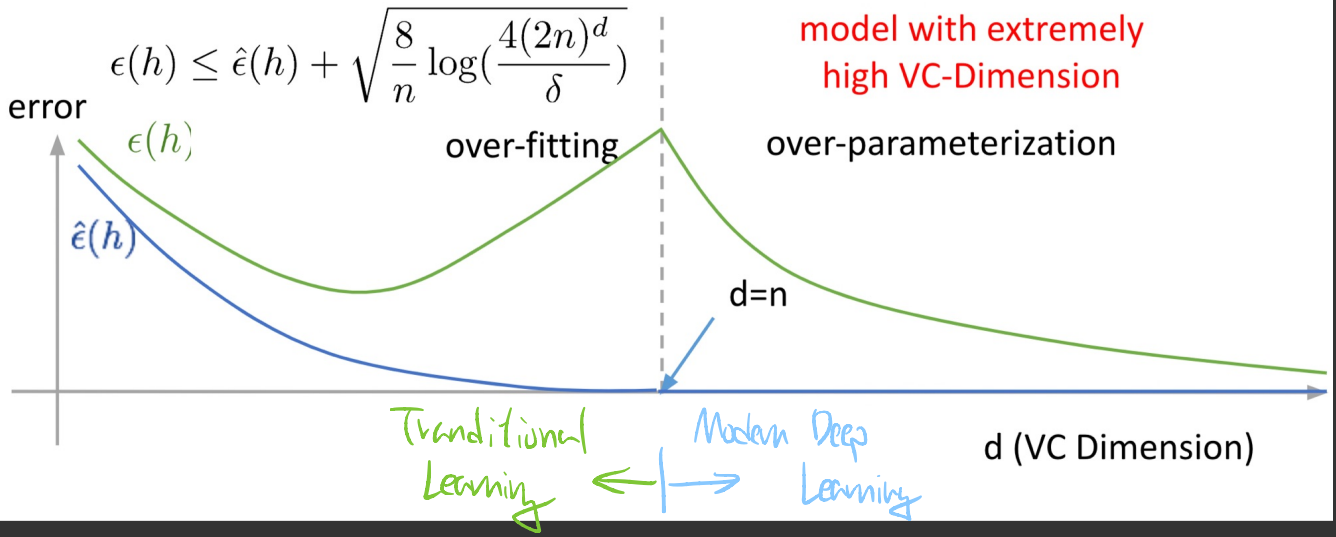
Better situation $\hat{\varepsilon}(h) \sim \varepsilon(h)$

$\hat{\varepsilon}(h)$: training error ; $\varepsilon(h)$: testing error

11/18. Fri. 2022

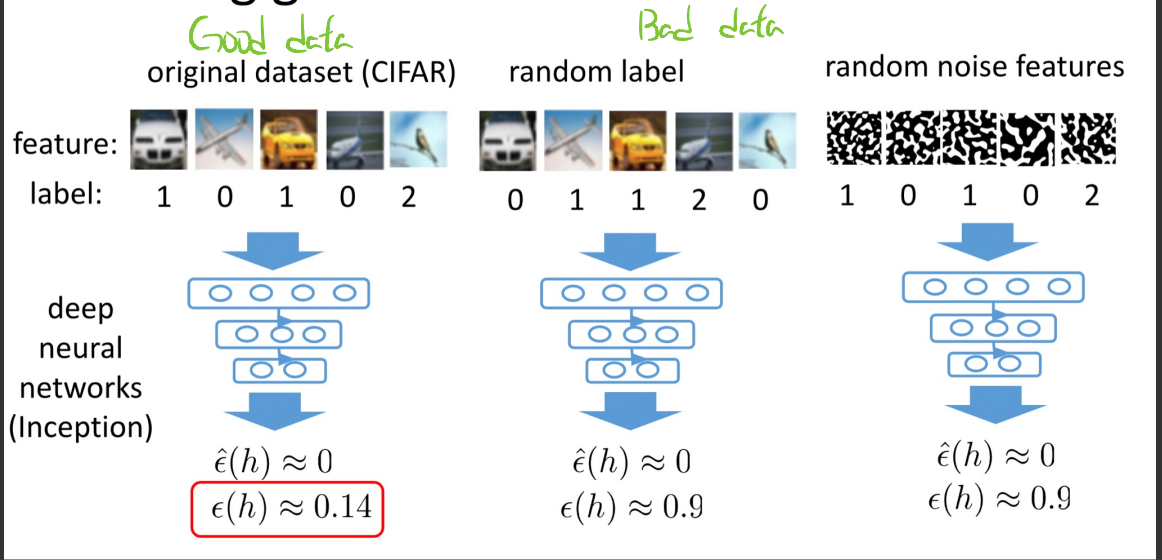
* Modern machine learning

Reconciling modern machine learning and the bias-variance trade-off



For modern model due to that hardware can handle very large data now,

Understanding deep learning requires rethinking generalization



基本上有資料都能練成一個模型，但資料牧說不佳會導致 testing 差勁

人工智慧是工人智慧

* Rethinking generalization for modern model

Regularization 對於不同模型效果不一，但改善 $\epsilon(h)$ 效果不大

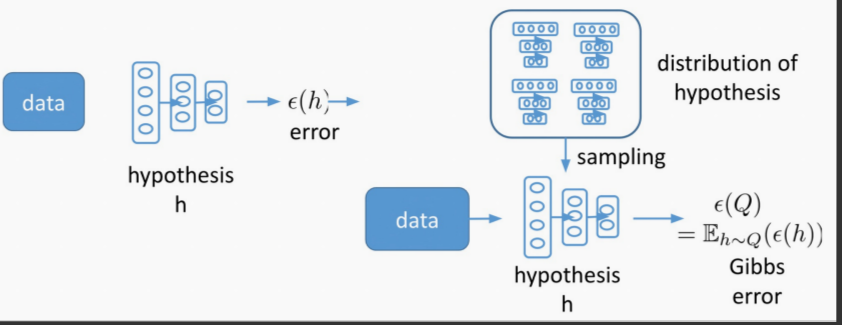
11/18. Fri. 2022

* Pac-Bayesian Framework.

Talking about how regularization affect models.

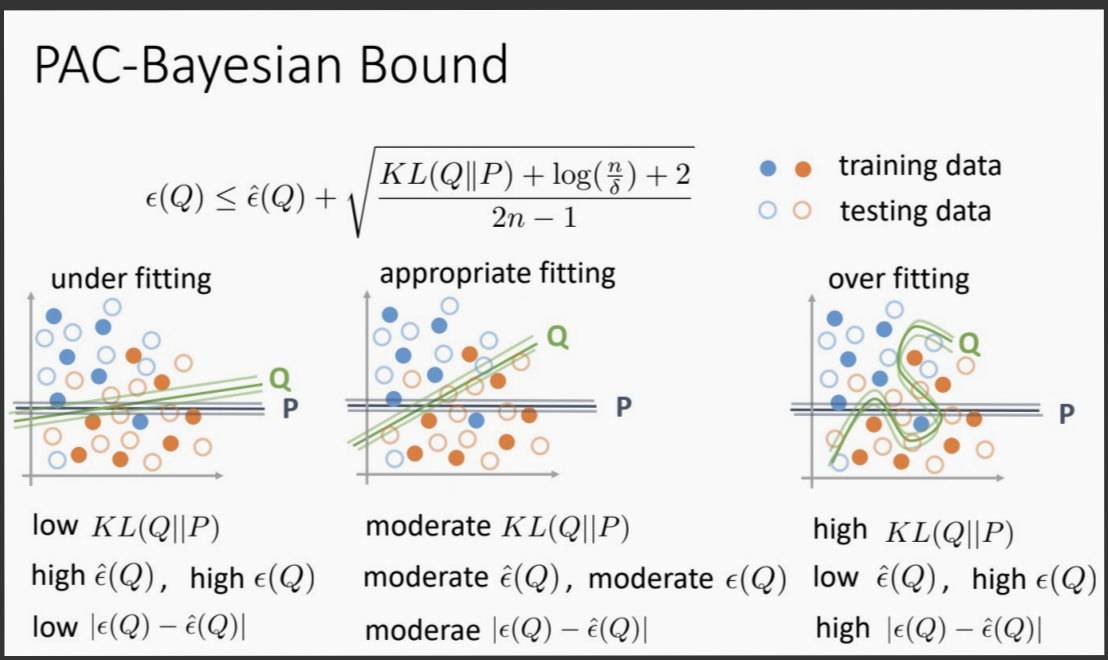
PAC-Bayesian Bound

- Deterministic Model
- Stochastic Model (Gibbs Classifier)



PAC-Bayesian Bound

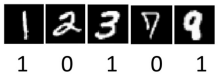
$$\epsilon(Q) \leq \hat{\epsilon}(Q) + \sqrt{\frac{KL(Q\|P) + \log(\frac{n}{\delta}) + 2}{2n-1}}$$



under fitting	appropriate fitting	over fitting
low $KL(Q\|\|P)$	moderate $KL(Q\|\|P)$	high $KL(Q\|\|P)$
high $\hat{\epsilon}(Q)$, high $\epsilon(Q)$	moderate $\hat{\epsilon}(Q)$, moderate $\epsilon(Q)$	low $\hat{\epsilon}(Q)$, high $\epsilon(Q)$
low $\lvert\epsilon(Q) - \hat{\epsilon}(Q)\rvert$	moderae $\lvert\epsilon(Q) - \hat{\epsilon}(Q)\rvert$	high $\lvert\epsilon(Q) - \hat{\epsilon}(Q)\rvert$

Stop training in a lower $|\epsilon(Q) - \hat{\epsilon}(Q)|$

Experiments



	original M-NIST	random label
Training	0.028	0.112
Testing	0.034	0.503
VC Bound	26m	26m
Pac-Bayesian Bound	0.161	1.352

Pac-Bayesian Bound 相對適合 Deep Learning.

11/18. Fri. 2022

* Comparison

How to Overcome Over-fitting?

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

Traditional Machine Learning

- reduce the number of parameters
- weight decay
- early stop
- data augmentation

資料優化

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

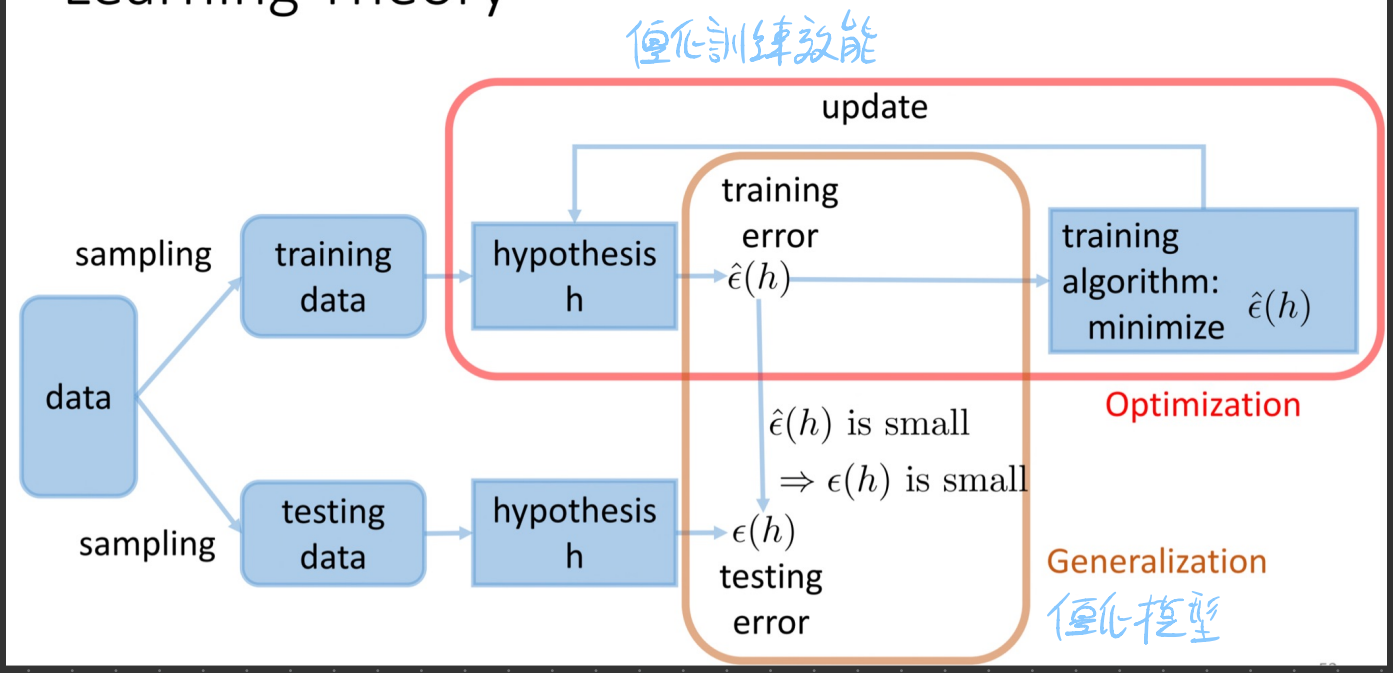
Modern Deep Learning Most Difference

- ~~reduce the number of parameters~~
- weight decay ?
- early stop ?
- data augmentation ?
- improving data quality
- starting from a good P

11/18. Fri. 2022

- Neural Tangent kernel

Learning Theory



Analysis of Generalization

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

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

    - Only depends on training data
    - Model doesn't need to be trained
    - Can only be applied to over-parameterized 2-layer ReLU NN
  - VC Bound

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

    - Only depends on model
    - Can not be applied to over-parameterization NN
  - PAC-Bayesian Bound

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

    - Depends on both model and training data
    - Model needs to be trained