Lecture 1

Dynamical Systems and Machine learning

Qianxiao Li
(NUS, A*STAR)

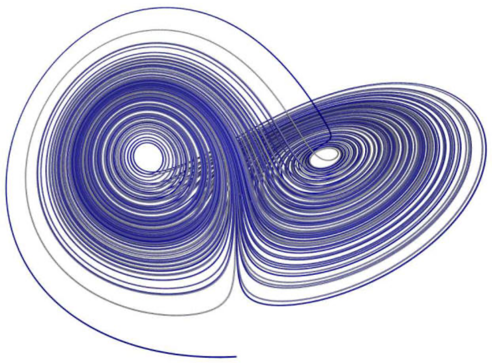
Overview

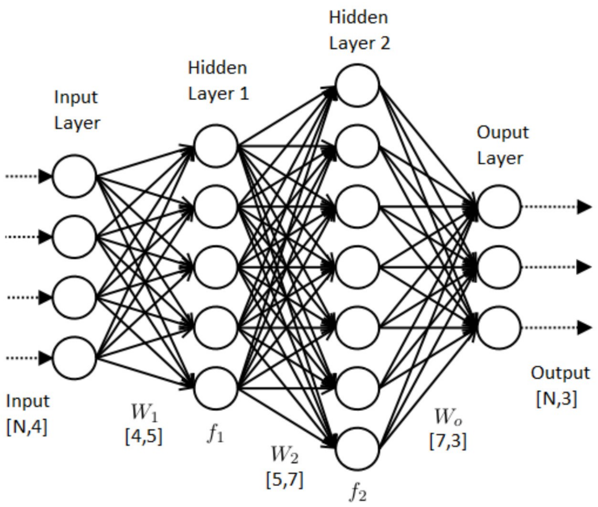
Dynamical Systems

Part I →

← Part II

Machine Learning





Syllabus

- Introduction
- Part I: Dynamical Systems Approach to Deep Learning
  - Optimal control theory
  - Optimal control formulation of deep learning
  - Dynamics/Control-inspired training algorithms
  - Dynamics/Control-inspired network architectures
  - Mathematical results
- Part II: Data-Driven Dynamical Systems
  - PCA/SVD based model reduction methods (DMD, POD)
  - Koopman operator methods
  - Data driven control

Other Matters

- Class participation is encouraged!
  - Use the chat
  - TA will moderate
- You can email me at qianxiao@nus.edu.sg
  - Questions about the material
  - Mistakes in lecture notes
  - Suggestions on lecture pace and style
  - Research opportunities

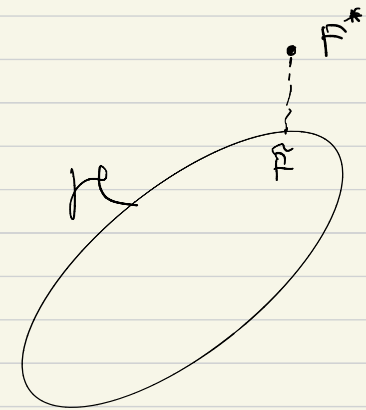
Supervised Learning

Basic formulation

- Data : $\mathcal{D} = \{x_i, y_i\}_{i=1}^N$
    - $x_i \in \mathbb{R}^d$ inputs
    - $y_i \in \mathbb{R}^m$ outputs / labels
    - $N \geq 1$ size of data
- Data distribution : $x_i \overset{\text{i.i.d.}}{\sim} \mu$.
- Target function :
    - Deterministic : $y_i = F^*(x_i)$
    - Stochastic : $y_i \sim p^*(\cdot \mid x_i)$ e.g. $y_i = F^*(x_i) + \varepsilon_i$ $N(0, I)$
    - General : $(x_i, y_i) \sim \mu$
- Output space
    - Regression $y_i \in \mathbb{R}$
    - Classification $y_i \in \{1, 2, \cdots, C\} \xrightarrow{\text{embed}} \mathbb{R}^C$ $\quad i \mapsto \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \\ 0 \\ \vdots \\ 0 \end{pmatrix}$

Goal of supervised learning

Given $\mathcal{D}$, find $\tilde{F} \approx F^*$



Learn with risk minimization

- Define a hypothesis space

$$\mathcal{H} = \{ F \mid F : \mathbb{R}^d \to \mathbb{R}^m \}$$

goal is to find $\tilde{F} \in \mathcal{H}$ s.t. $\| F^* - \tilde{F} \|$ is small.

- Define a loss function

$$\Phi : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R}.$$

- Define the empirical risk

$$R_{emp}(F) = \frac{1}{N} \sum_{i=1}^{N} \Phi(F(x_i), y_i)$$

$y_i = F^*(x_i)$

- Empirical risk minimization

$$\hat{F} \leftarrow \underset{F \in \mathcal{H}}{\arg\min}\, R_{emp}(F)$$

- Loss function
  - mean-square loss

$$\Phi(y, y') = \frac{1}{2}\|y - y'\|^2$$

  - zero-one loss

$$\Phi(y, y') = \mathbb{1}_{y \neq y'} = \begin{cases} 1 & y \neq y' \\ 0 & y = y' \end{cases}$$

- Population/Expected Risk Minization (PRM)

$$\underset{F \in \mathcal{H}}{\operatorname{argmin}} \; \mathbb{E}_{x \sim \mu} \, \Phi(F(x), \underbrace{F^*(x)}_{y}) \qquad \text{(deterministic)}$$

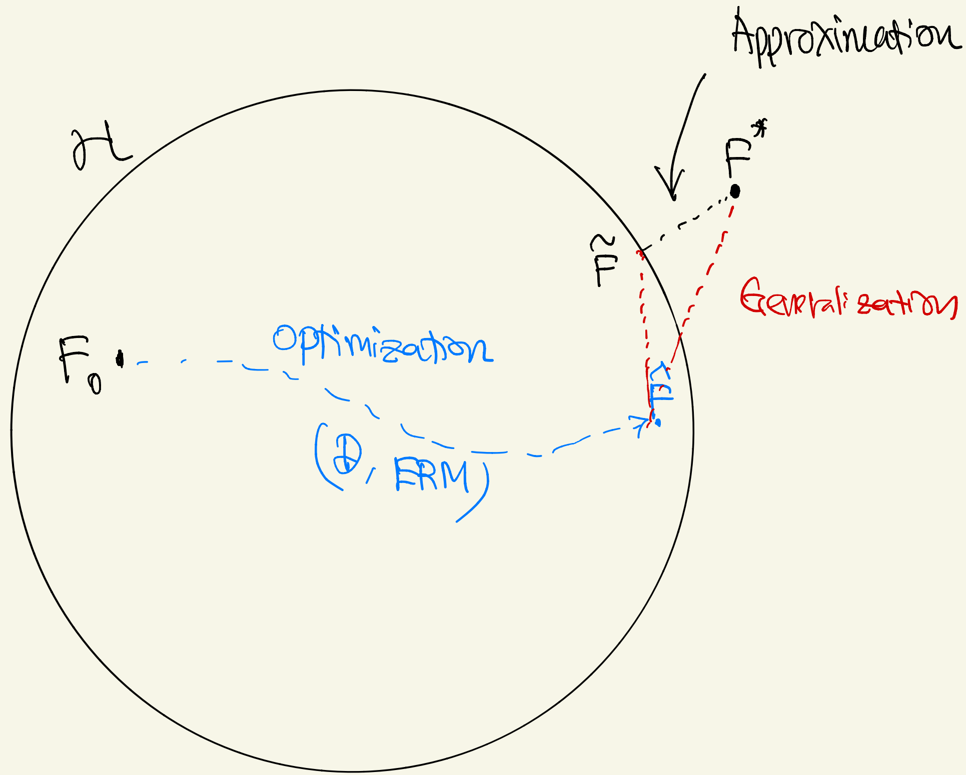
$$\tilde{F} \longleftarrow \underset{F \in \mathcal{H}}{\operatorname{argmin}} \; \mathbb{E}_{(x,y) \sim \mu} \, \Phi(F(x), y) \qquad \text{(general)}$$

If μ is empirical measure

$$\mu = \frac{1}{N} \sum_{i=1}^{N} \delta_{(x_i, y_i)}$$

ERM.

Three paradigms of Supervised Learning



Example (Linear Models)

Regression $y \in \mathbb{R}$

Linear model hypothesis space

$$\mathcal{H} = \left\{ F : F(x) = \sum_{j=0}^{M-1} w_j \phi_j(x) \ , \ w_j \in \mathbb{R}, \ j = 0, 1, \cdots, M-1 \right\}$$

Basis fns, feature maps.

$\phi_j : \mathbb{R}^d \to \mathbb{R}$.

Examples of Basis Function $(d=1)$

- $\phi_0(x) = 1$, $\phi_1(x) = x \Rightarrow F(x) = w_0 + w_1 x$.
- Polynomial : $\phi_j(x) = x^j$
- RBF / Gaussian : $\phi_j(x) = \exp\left(-\frac{1}{2s_j^2}(x - m_j)^2\right)$

- Empirical Risk Minimization

$$\frac{1}{2}(f(x_i)-y_i)^2$$

$$R_{emp}(\omega_0,\dots,\omega_{M-1}) = \frac{1}{N}\sum_{i=1}^{N}\Phi(f(x_i),y_i)$$

$$= \frac{1}{2N}\sum_{i=1}^{N}\Big(\underbrace{\sum_{j=0}^{M-1}\omega_j\phi_j(x_i)}_{f(x)} - y_i\Big)^2$$

Matrix form

$$R_{emp}(\omega) = \frac{1}{2N}\|\Phi\omega - y\|^2$$

$$\omega = \begin{pmatrix}\omega_0\\ \vdots \\ \omega_{M-1}\end{pmatrix}\in\mathbb{R}^M,\quad y=\begin{pmatrix}y_1\\ \vdots \\ y_N\end{pmatrix}\in\mathbb{R}^N,\quad (\Phi)_{ij} = \phi_j(x_i)\in\mathbb{R}^{N\times M}.$$

$$\min_{\omega} R_{emp}(\omega) \rightarrow \hat{\omega}$$

Ordinary Least Squares Formula:

Suppose $\Phi^T\Phi$ is invertible, then

$$\hat{w} = (\Phi^T\Phi)^{-1}\Phi^T y.$$

Proof

$$R_{emp}(w) = \frac{1}{2N}\|\Phi w - y\|^2$$

$$\nabla_w R_{emp}(w) = \frac{1}{N}\Phi^T(\Phi w - y)$$

Set $\nabla_w R_{emp}(\hat{w}) = 0$

$$\Rightarrow (\Phi^T\Phi)\hat{w} = \Phi^T y.$$

$$\hat{w} = (\Phi^T\Phi)^{-1}\Phi^T y. \quad \square$$

If $M > N$, $\Phi^T\Phi$ not invertible. $(\hat{w} = \Phi^\dagger y)$ $\Phi^\dagger \in \mathbb{R}^{M\times N}$.

Consider regularized problem

$$\min_w \frac{1}{2N}\|\Phi w - y\|^2 + \lambda\|w\|^2 \quad \Rightarrow \quad \hat{w} = (\lambda I + \Phi^T\Phi)^{-1}\Phi^T y.$$

$(\lambda > 0)$

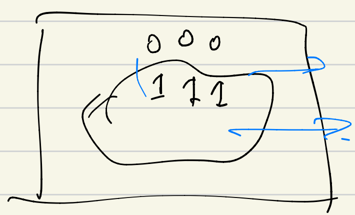
$F^*: \mathbb{R}^d \to \mathbb{R}^m$.

$\mathcal{G} \to$ group of transformations. $\quad g(x)$ group action.

$$F^*(g(x)) = F^*(x) \quad \text{(invariant)}$$

(equivariant) $$F^*(g(x)) = \hat{g}(F^*(x))$$ $\quad \hat{g}(x)$ action on $\mathbb{R}^m$.

$\mathcal{H}$



$$R_{pop}(\hat{F}) - R_{emp}(\hat{F}) \lesssim \frac{C(\hat{F})}{N^{\alpha}}.$$

Neural Network Hypothesis Space

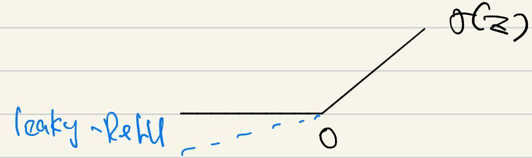
$$\mathcal{H}_M = \left\{ F : F(x) = \sum_{j=1}^{M} v_j \underbrace{\sigma(w_j^T x + b_j)}_{\phi_j(x)} ,\ w_j \in \mathbb{R}^d, v_j, b_j \in \mathbb{R}, j=1,\dots,M \right\}$$

- "adaptive basis model"
- σ is the activation function

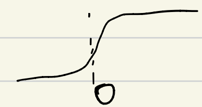
  $\sigma : \mathbb{R} \to \mathbb{R}$.

  Examples :

  - ReLU (Rectified linear unit) $\sigma(z) = \max(0, z)$



  - Sigmoid $\sigma(z) = \dfrac{1}{1+e^{-z}}$



  - tanh

- Matrix form

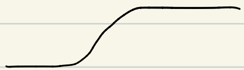
$$F(x) = v^T \sigma(Wx + b) \qquad W \in \mathbb{R}^{M \times d}.$$

$$[\sigma(z)]_i = \sigma(z_i)$$

Universal Approximation Theorem

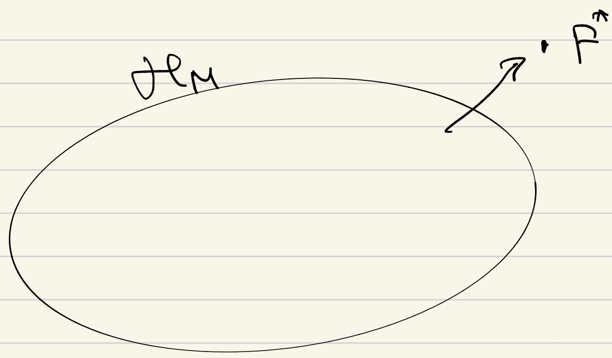
Let $K \subset \mathbb{R}^d$ be compact; $F^*: K \to \mathbb{R}$ be continuous.

Assume $\sigma: \mathbb{R} \to \mathbb{R}$ is sigmoidal, i.e. it is continuous and $\lim_{z \to -\infty} \sigma(z) = 0$ $\lim_{z \to +\infty} \sigma(z) = 1$



Then, $\forall \varepsilon > 0$, $\exists F \in \cup_M \mathcal{H}_M$ s.t.

$$\|F^* - F\|_{C(K)} = \sup_{x \in K} |F^*(x) - F(x)| \leq \varepsilon.$$



Optimizing / Training NN

Euclidean space.

- $\mathcal{H} = \{ F_\theta : \theta \in \Theta \}$
- ERM $\min_{\theta \in \Theta} R(\theta) \leftarrow \min_{F \in \mathcal{H}} R(F)$
- If R is C^1

  necessary condition for optimality is

  $$\nabla_\theta R(\hat{\theta}) = 0$$

- Gradient descent:

  $$\theta_{k+1} = \theta_k - \eta \nabla R(\theta_k) \qquad k = 0, 1, \dots$$

  $\eta > 0$ is the learning rate.

Suppose $\theta_k \to \theta_\infty$

$$\cancel{\theta_\infty} = \cancel{\theta_\infty} - \underbrace{\eta \nabla R(\theta_\infty)}_{=0}$$

- Local vs Global Minima.

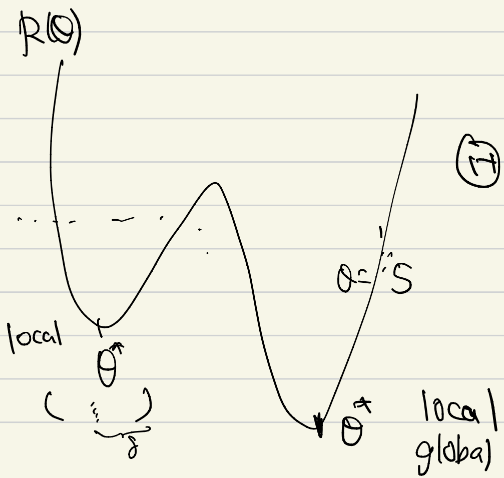
$\theta^* \in \Theta$ is a local minimum of R if

$\exists\ \delta > 0$ s.t.

$$R(\theta^*) \le R(\theta) \quad \forall\ \|\theta - \theta^*\| \le \delta .$$

〃 〃 〃 global 〃 〃

$$R(\theta^*) \le R(\theta) \quad \forall\ \theta \in \Theta$$



$\Theta = \mathbb{R}.$

$\Theta = [5, \infty)$

The notion of local/global minima depends on

- Θ
- $\|\cdot\|$

Deep Neural Networks

$$\mathcal{H} = \{ F : \quad F(x) = v^T x(K), \quad v \in \mathbb{R}^{d_K}$$

where $\quad x(k+1) = \sigma(W(k)x(k) + b(k))$

$k = 0, 1, \dots, K-1$

$x(0) = x.$ (input)

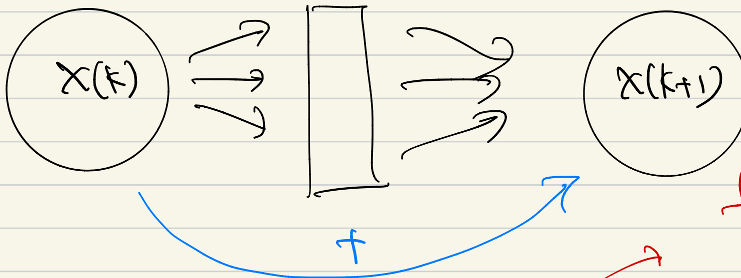
$W(k) \in \mathbb{R}^{d_{k+1} \times d_k}, \quad b(k) \in \mathbb{R}^{d_{k+1}} \quad \}$

input

$$x = x(0) \longrightarrow \underbrace{\boxed{x \mapsto \sigma(W(0)x + b(0))} \longrightarrow x(1) \longrightarrow \boxed{x \mapsto \sigma(W(1)x + b(1))} \longrightarrow x(2) \longrightarrow \cdots \longrightarrow x(K)}_{K \text{ layers}} \xrightarrow{v^T} \hat{y}.$$

Residual NN

$$x(k+1) = x(k) + \sigma(W(k)x(k) + b(k))$$



$f(k, x, \theta) = \sigma(Wx + b)$

Generally, we can write.

hidden state

weights / trainable params.

$$x(k+1) = x(k) + f(k, x(k), \theta(k))$$

layer id.

$$k = 0, 1, 2, \cdots, K-1$$

\# of layers or depth

$$f : \mathbb{Z}^{+} \times \mathbb{R}^{d} \times \Theta \to \mathbb{R}^{d}.$$