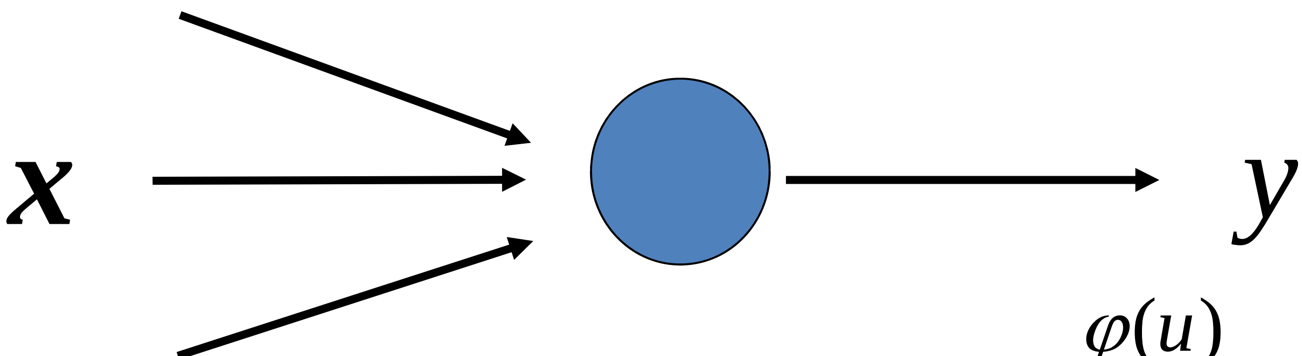
特異モデルと
学習のダイナミクス

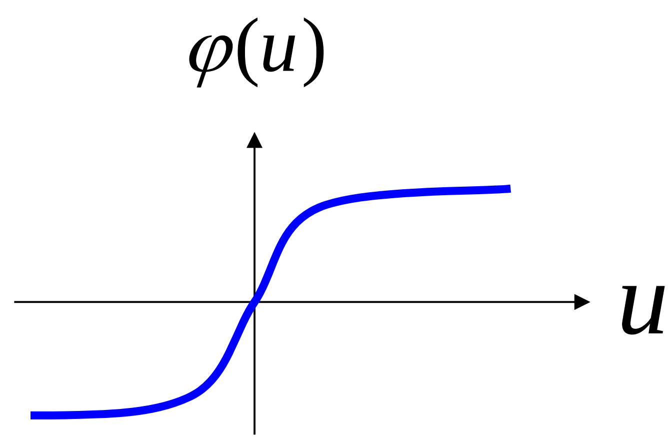
甘利俊一　理研脳科学総合研究センター

尾関智子、Florent Cousseau, Hyeyoung Park

Mathematical Neurons

$$y = \varphi\left(\sum w_i x_i - h\right) = \varphi\left(\boldsymbol{w} \cdot \boldsymbol{x}\right)$$

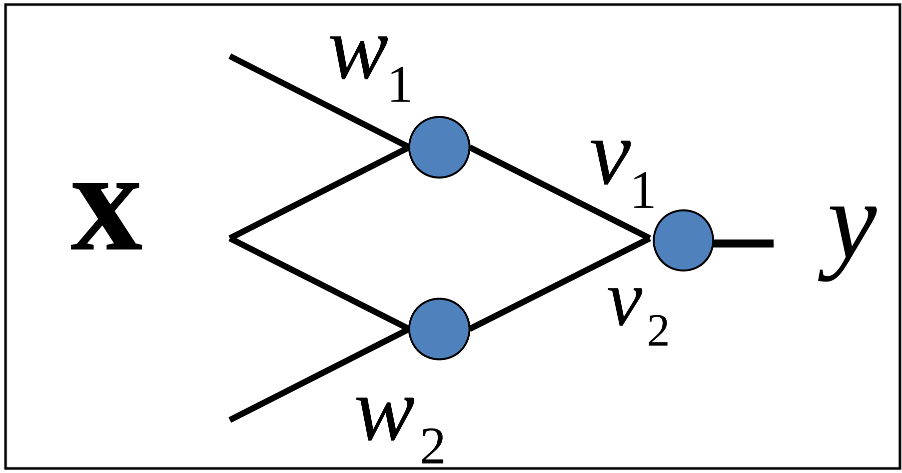




Multilayer Perceptrons

$$y = \sum v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right) + n$$

$$x = \left(x_1, x_2, \ldots, x_n\right)$$



$$p\left(y \middle| \boldsymbol{x}; \boldsymbol{\theta}\right) = c \exp\left\{-\frac{1}{2}\left(y - f\left(\boldsymbol{x}, \boldsymbol{\theta}\right)\right)^2\right\}$$

$$f\left(\boldsymbol{x}, \boldsymbol{\theta}\right) = \sum v_i \varphi\left(\boldsymbol{w}_i \cdot \boldsymbol{x}\right)$$

$$\theta = \left(w_1, \ldots, w_m; v_1, \ldots, v_m\right)$$

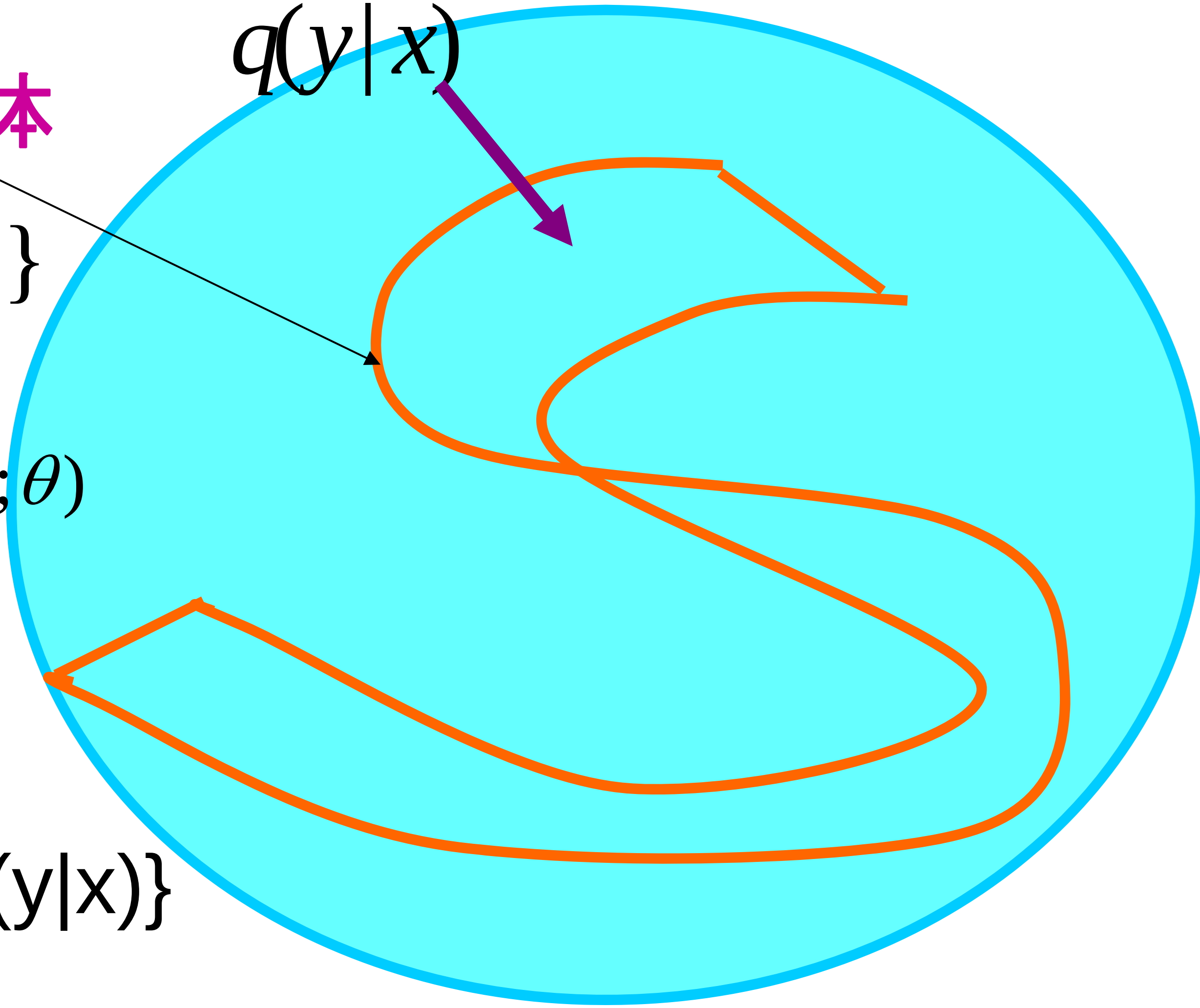
Multilayer Perceptron

$q(y|x)$

神経多様体

$$M = \{ p(y \mid x; \theta) \}$$

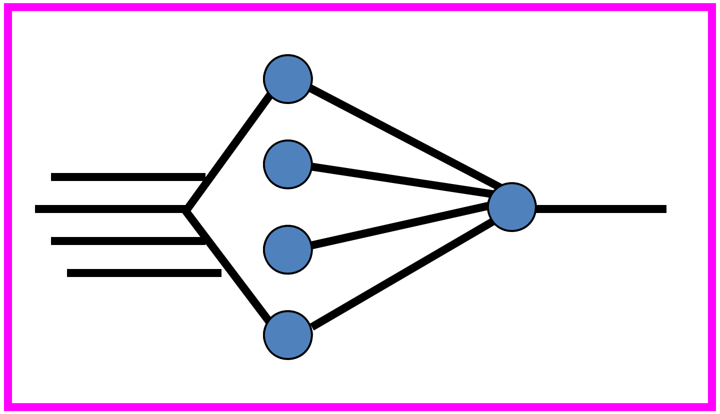
$$p(y, x; \theta) = q(x)\, p(y \mid x; \theta)$$



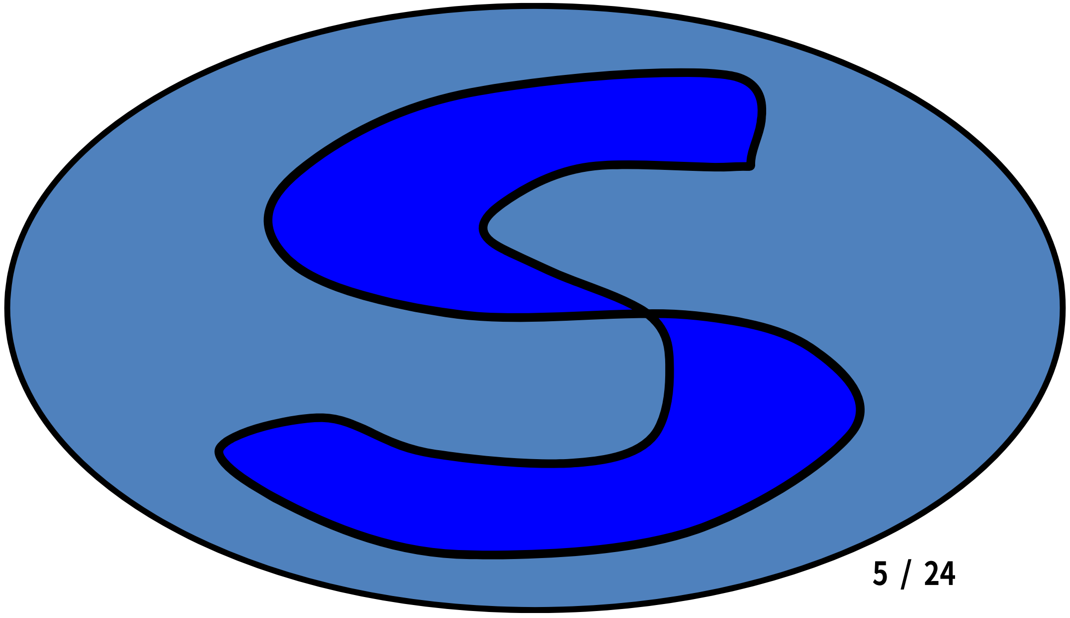
space of {q(y|x)}

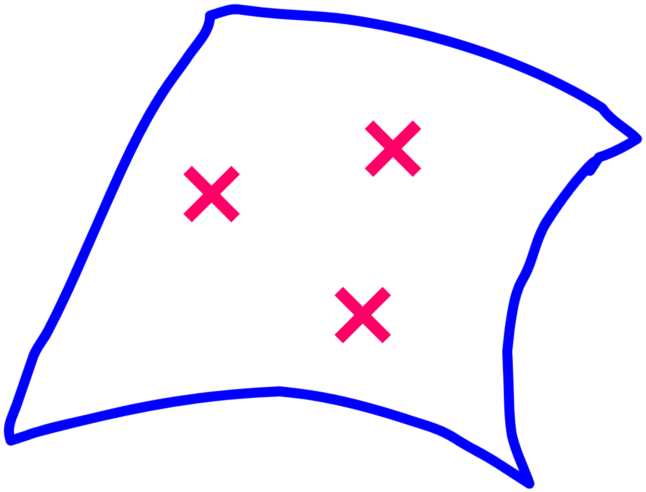
神経多様体

- 計量構造
- 位相構造

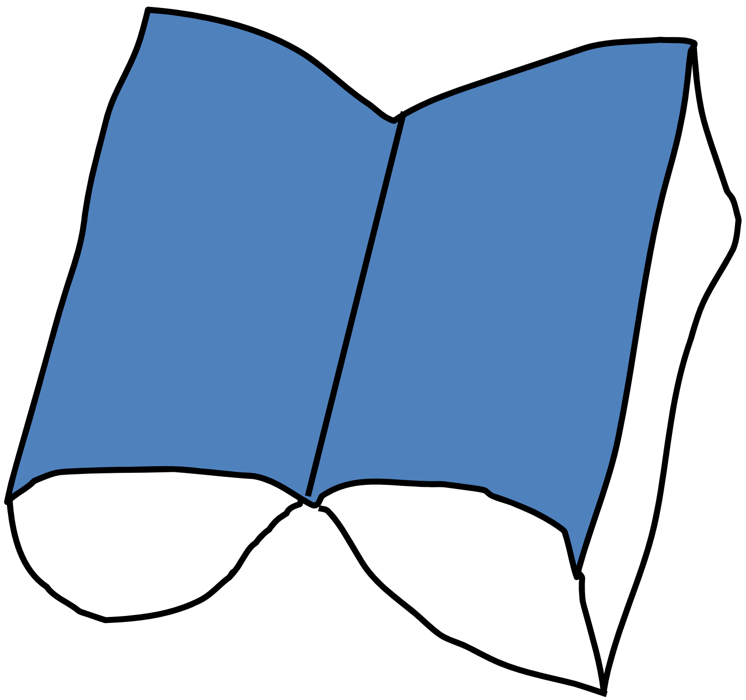


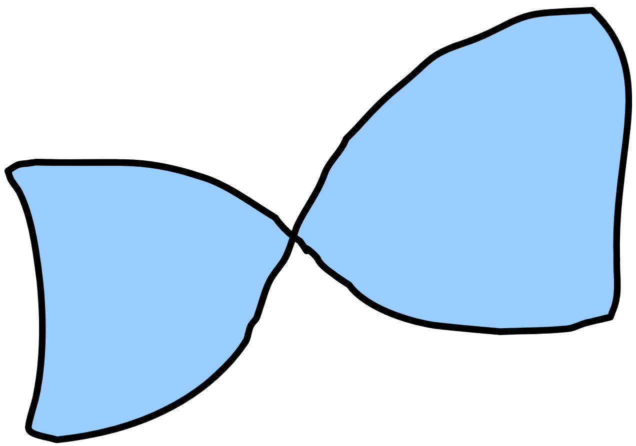
θ





singularities—特異点



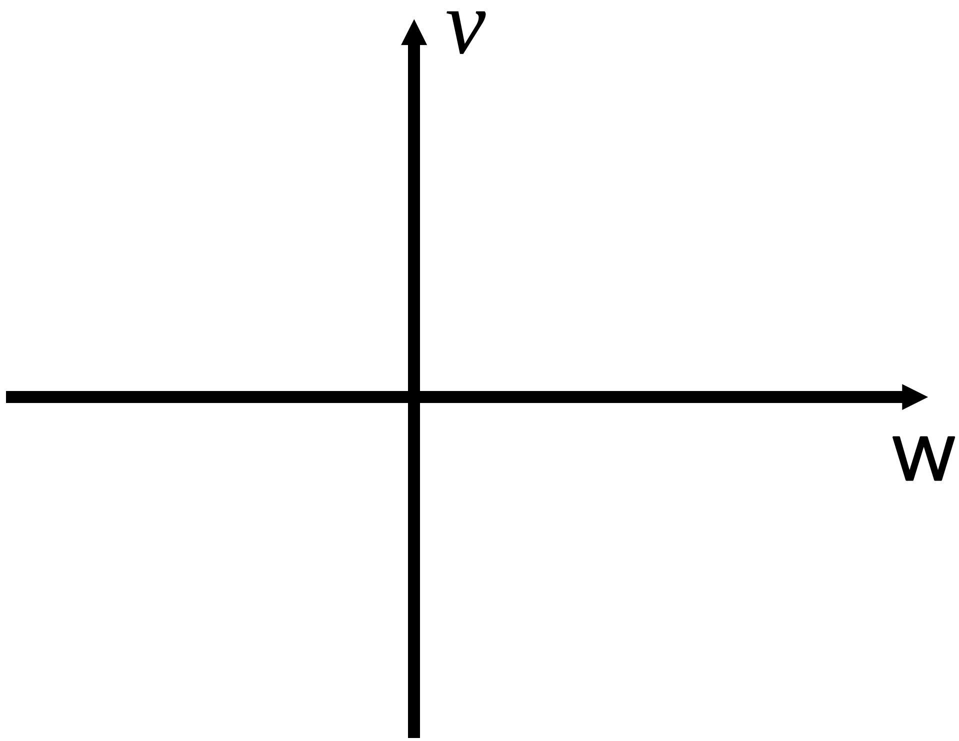


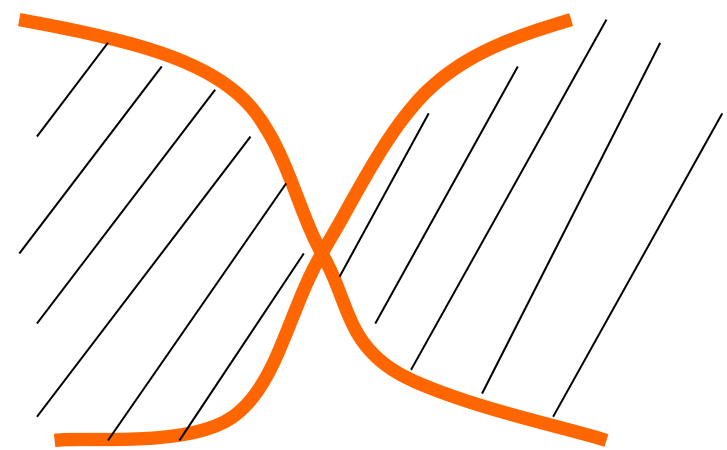
微分幾何—代数幾何

Geometry of singular model

$$y = v\varphi(\boldsymbol{w}\cdot\boldsymbol{x}) + n$$

$$v\,|\,\mathbf{w}\,| = 0$$



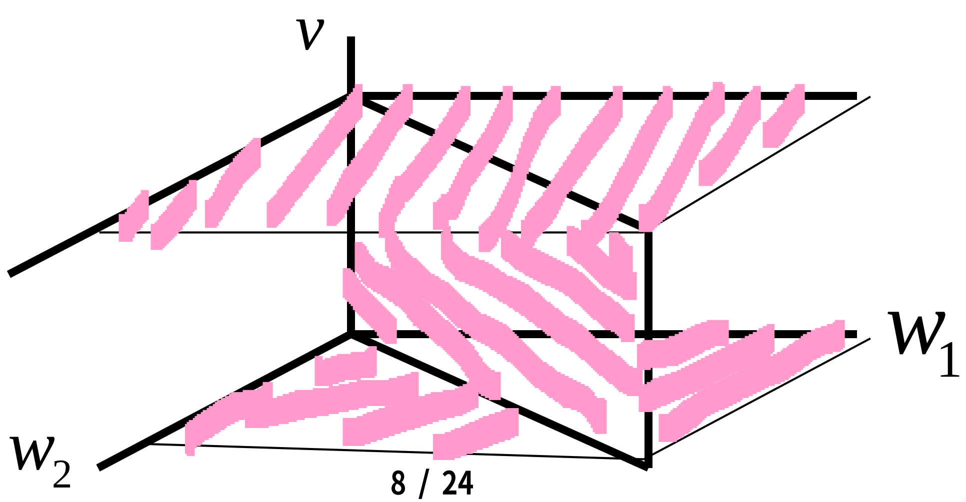


Gaussian mixture

$$p(x; v, w_1, w_2) = (1-v)\varphi(x-w_1) + v\varphi(x-w_2)$$

$$\varphi(x) = \frac{1}{\sqrt{2\pi}} \exp\left\{-\frac{1}{2}x^2\right\}$$

singular: $w_1 = w_2$, $v(1-v) = 0$



特異モデルの解析

1. 真の分布が特異点にあるとき
2. 真の分布がそれ以外のとき

1. 推論一尤度比の奇妙な振舞い一福水
2. 推定量の振舞い
3. Bayes 推定量一渡辺ら
4. 学習ダイナミックス

Regular statistical model

$$M = \left\{ p\left(x, \theta\right) \right\}$$

G : Fisher information

$$E\left[\Delta\theta \Delta\theta^{T} \right] = \frac{1}{n} G^{-1}$$

$$E\left[KL\left[p\left(x, \theta_0\right) : p\left(x, \hat{\theta}\right) \right] \right] \approx \frac{1}{2n} G \cdot E\left[\Delta\theta \Delta\theta \right]$$

$$\approx \frac{d}{2n}$$

AIC, BIC, MDL

$$\lambda = 2\sum \log \frac{p(y_i, \boldsymbol{x}_i, \hat{\theta})}{p(y_i, \boldsymbol{x}_i, \theta_0)}$$

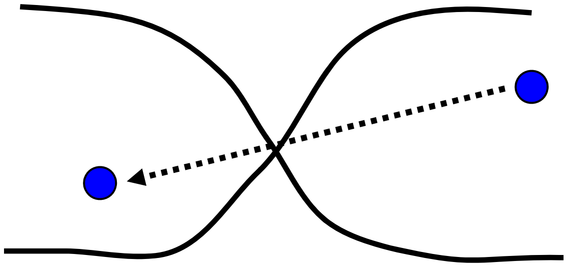
$$\lambda = n(\hat{\theta} - \theta_0)^T G^{-1}(\theta_0)(\hat{\theta} - \theta_0)$$

$$E\left[\lambda\right] = c(n)k$$

$$c(n) = \log n$$

$$c(n) = \sqrt{\log \log n}$$

Learning, Estimation, and Model Selection



$$E_{\text{gen}} = D\left[p_0\left(y \middle| \boldsymbol{x}\right) : p\left(y \middle| \boldsymbol{x};\hat{\boldsymbol{\theta}}\right)\right]$$

$$E_{\text{train}} = D\left[p_{\text{emp}}\left(y \middle| \boldsymbol{x};\hat{\boldsymbol{\theta}}\right)\right]$$

$$E_{\text{gen}} = \frac{d}{2n} \qquad d : \text{dimension}$$

$$E_{\text{gen}} = E_{\text{train}} + \frac{d}{n}$$

Learning from examples

$$\psi\left(\boldsymbol{x}\right) \approx f\left(x, \hat{\theta}\right) = \sum v_i \phi(\mathbf{w}_i \bullet \mathbf{x}_i)$$

Training set T

$$\text{examples} \cdots \left(\boldsymbol{x}_1, y_1\right), \cdots, \left(\boldsymbol{x}_n, y_n\right)$$

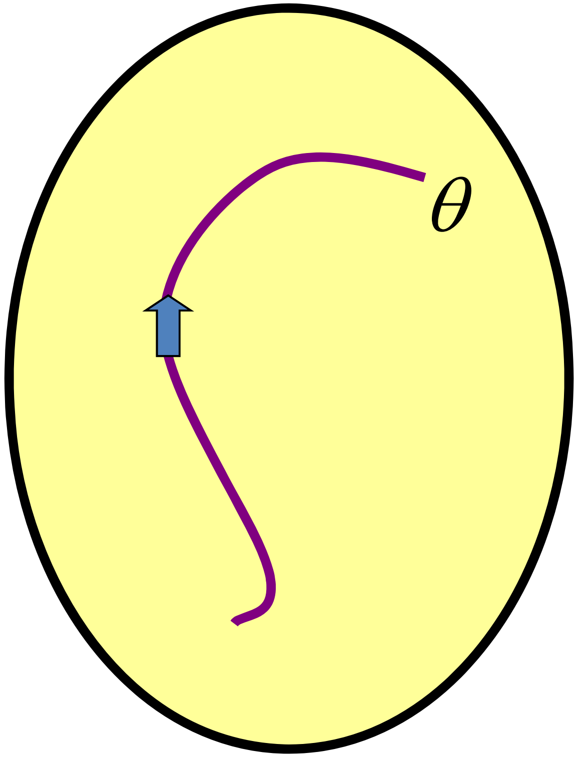
learning ; estimation

Backpropagation ---gradient learning

$$\text{examples}:\left(y_1,\boldsymbol{x}_1\right),\cdots\left(y_t,\boldsymbol{x}_t\right)--\text{training set}$$

$$l(y,x;\theta)=\frac{1}{2}\left|y-f\left(\boldsymbol{x},\boldsymbol{\theta}\right)\right|^2$$

$$=-\log p\left(y,\boldsymbol{x};\boldsymbol{\theta}\right)$$



$$\Delta\boldsymbol{\theta}_t=-\eta_t\frac{\partial l}{\partial\boldsymbol{\theta}}$$

$$f\left(\boldsymbol{x},\boldsymbol{\theta}\right)=\sum v_i\varphi\left(\boldsymbol{w}_i\cdot\boldsymbol{x}\right)$$

Problem of Backprop

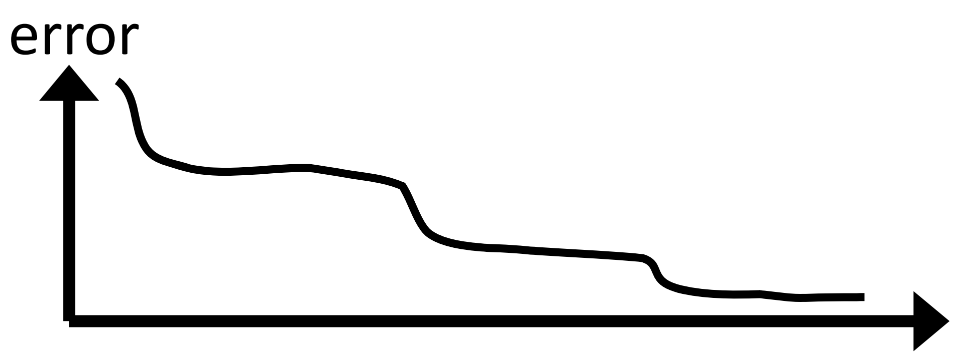
- **slow convergence----plateau---saddle**

- **local minima**

$$\Delta\theta_t = -\eta_t \nabla l(x_t, y_t; \theta_t)$$

Flaws of MLP

slow convergence : Plateau



local minima

→ Boosting and Bagging

Natural Gradient

$$\max \qquad dl = l\left(\theta + d\theta\right) - l\left(\theta\right)$$

$$\left|d\theta\right|^2 = \varepsilon$$

$$\tilde{\nabla} l = G^{-1}\left(\theta\right)\nabla l$$

$$\Delta\theta_t = -\eta_t \tilde{\nabla} l(x_t, y_t; \theta_t)$$

Information Geometry of MLP

Natural Gradient Learning :
S. Amari ; H.Y. Park

$$\Delta \boldsymbol{\theta} = -\eta G^{-1}(\boldsymbol{\theta}) \frac{\partial l}{\partial \boldsymbol{\theta}}$$

$$G_{t+1}^{-1} = (1+\varepsilon) G_t^{-1} - \varepsilon G_t^{-1} \nabla f \nabla f^T G_t^{-1}$$

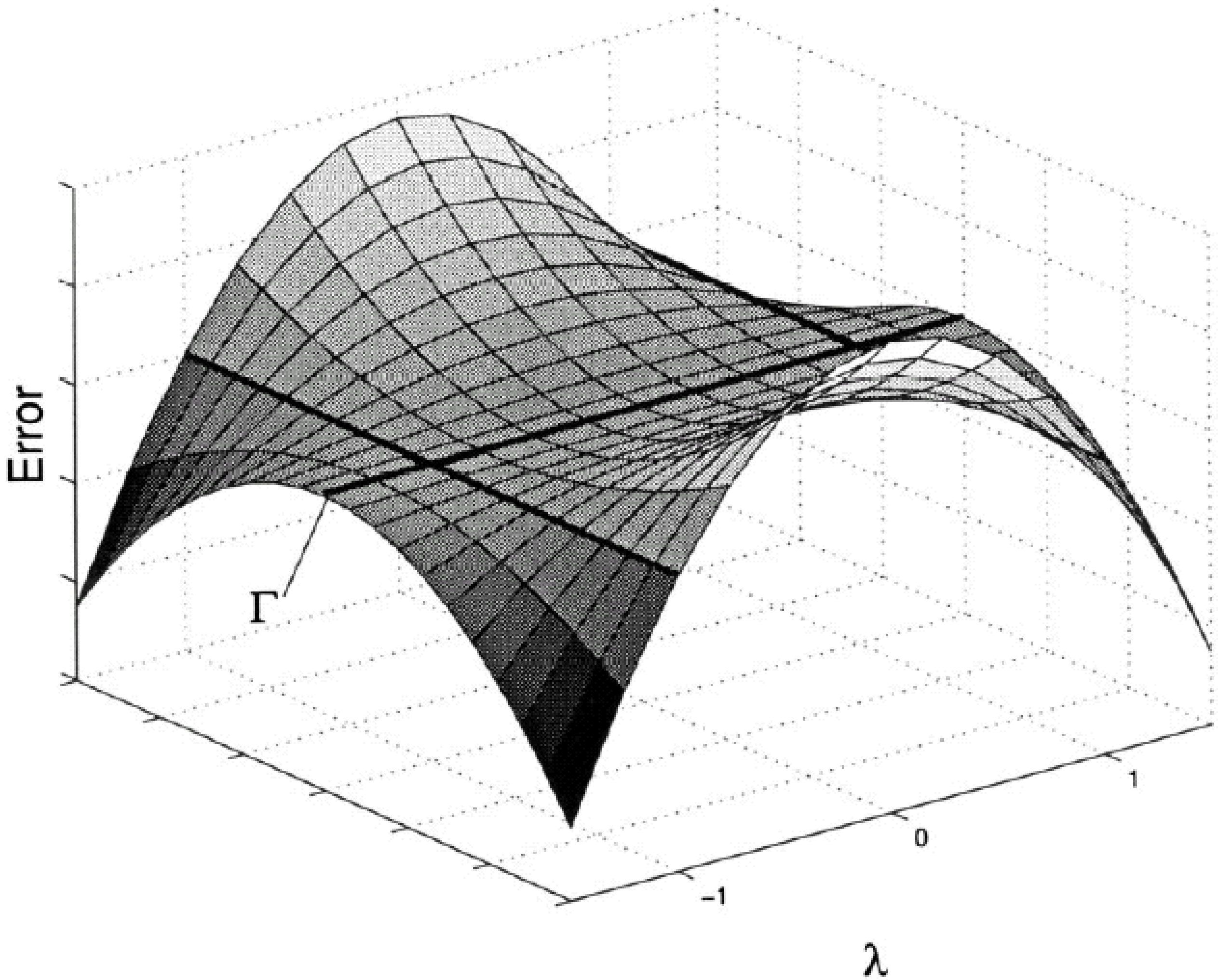
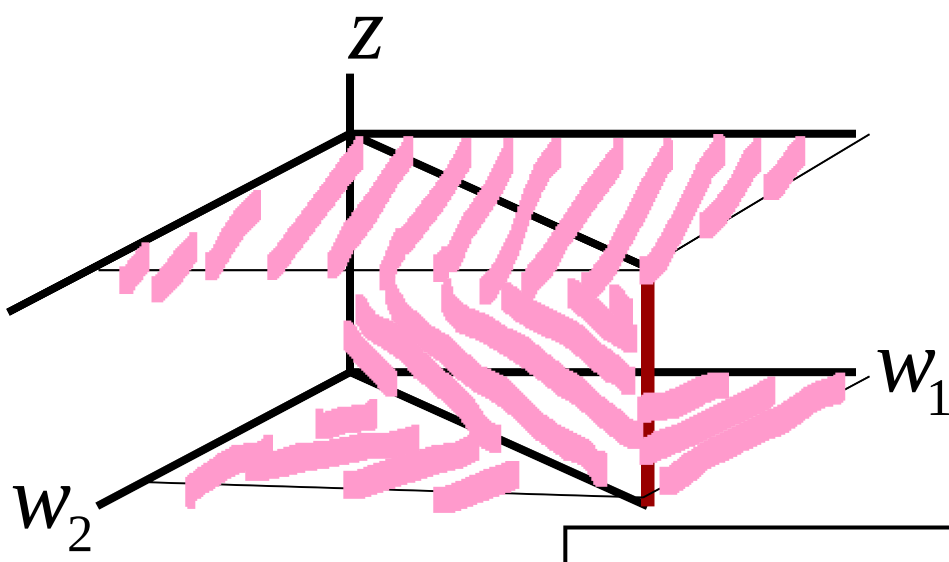


Fig. 5. Critical set with local minima and plateaus.

$$y = v_1 \varphi(w_1 x) + v_2 \varphi(w_2 x) + n$$

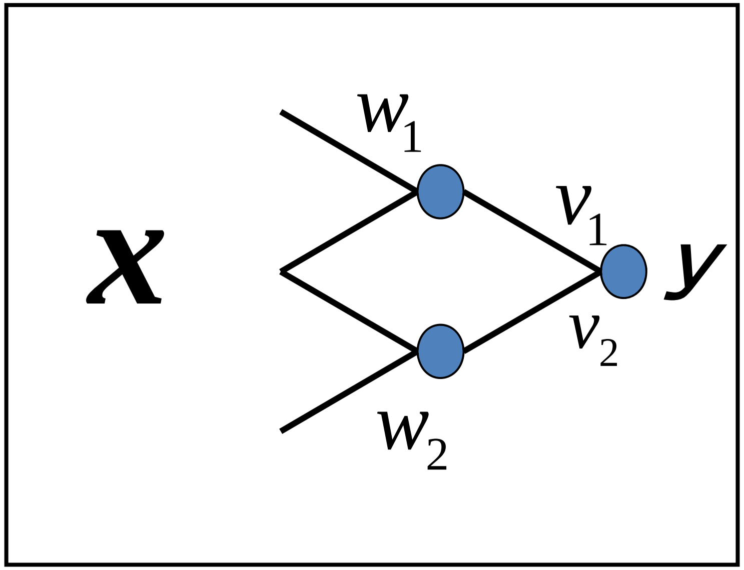
$$w_1 = w_2 = w$$

$$v_1 + v_2 = v$$



$$u = w_2 - w_1$$

$$z = v_2 - v_1$$



Coordinate Transformation

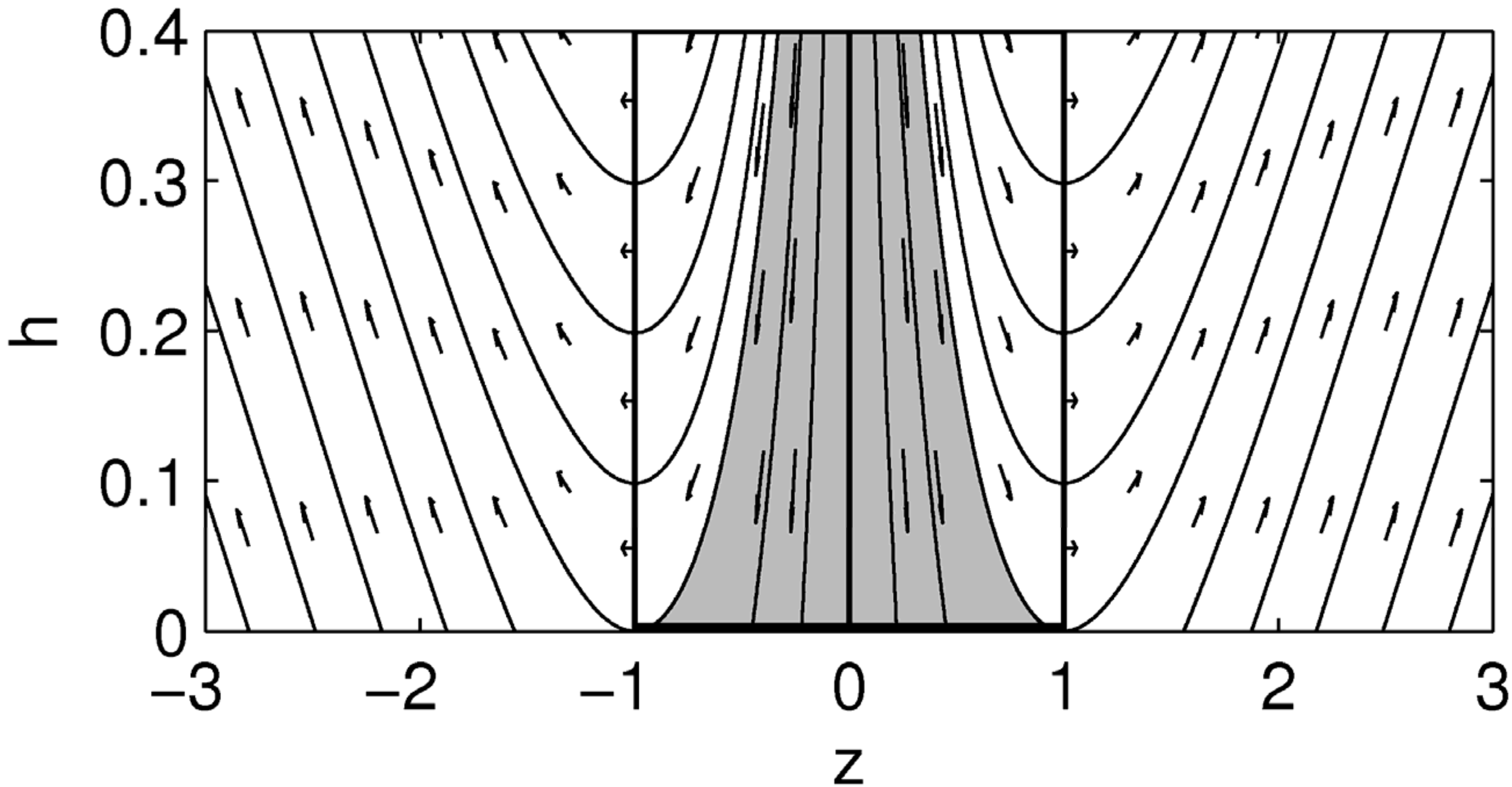
$$
\begin{cases}
\boldsymbol{u} = \boldsymbol{w}_2 - \boldsymbol{w}_1 & : \boldsymbol{u} = 0 \qquad \mathcal{R}_1 \\
\boldsymbol{w} = \dfrac{v_1 \boldsymbol{w}_1 + v_2 \boldsymbol{w}_2}{v} & \boldsymbol{w} = \boldsymbol{w}^* \\
v = v_1 + v_2 & v = v^* \\
z = \dfrac{v_2 - v_2}{v} & z = \pm 1 \qquad \mathcal{R}_2
\end{cases}
$$

学習方程式

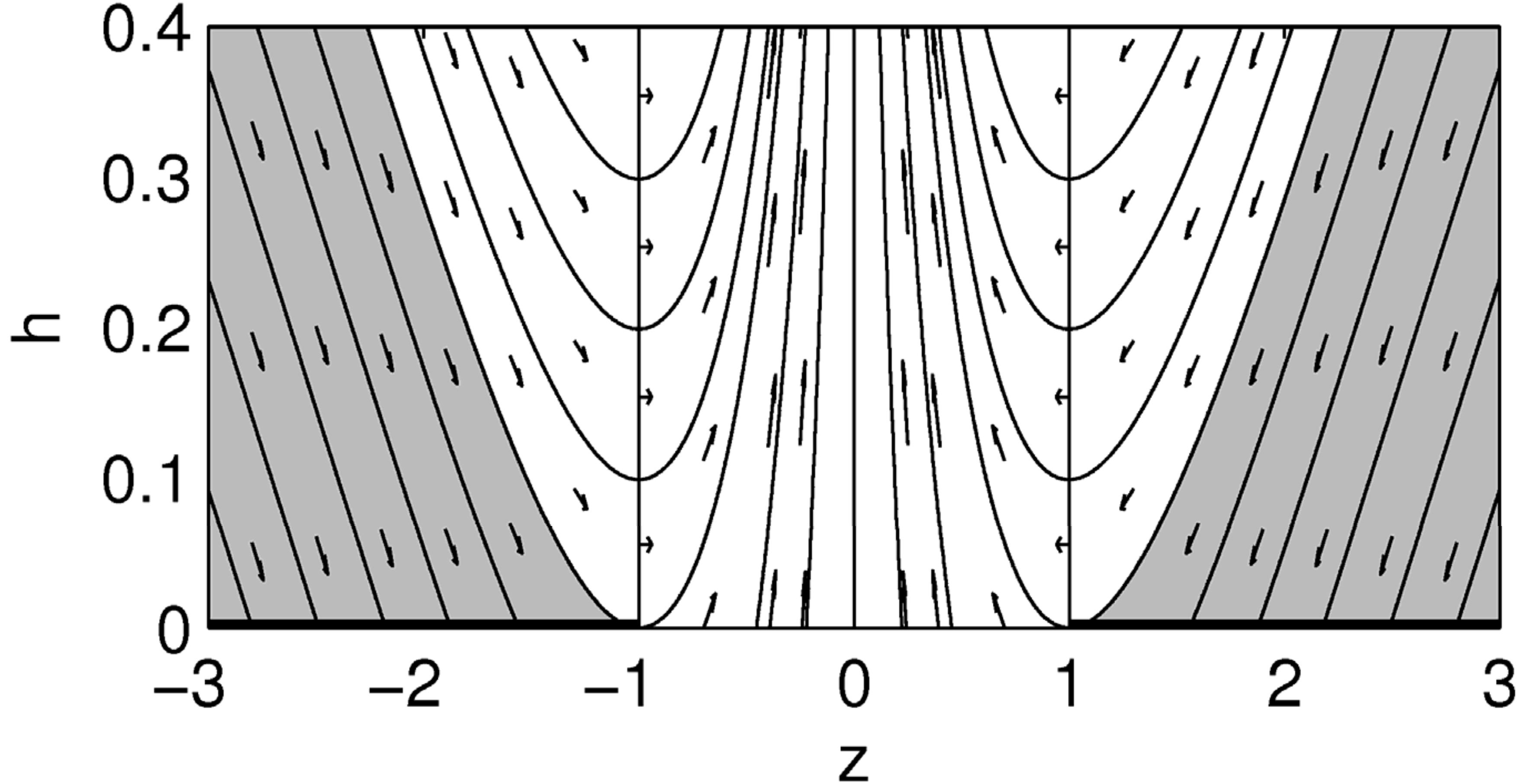
$$\frac{d\theta}{dt} = -\eta \nabla l, \qquad \frac{d\theta}{dt} = -\eta G^{-1} \nabla l$$

$$\frac{du}{dt} = f(u,z), \qquad \frac{dz}{dt} = k(u,z)$$

$$\frac{du}{dz} = \frac{f(u,z)}{k(u,z)}, \qquad u^2 = z^2 - \frac{1}{2}\log|z| + c$$



Dynamic vector fields: General case (|z|<1 part stable)



Dynamic vector fields: General case ($|z|>1$ part stable)