

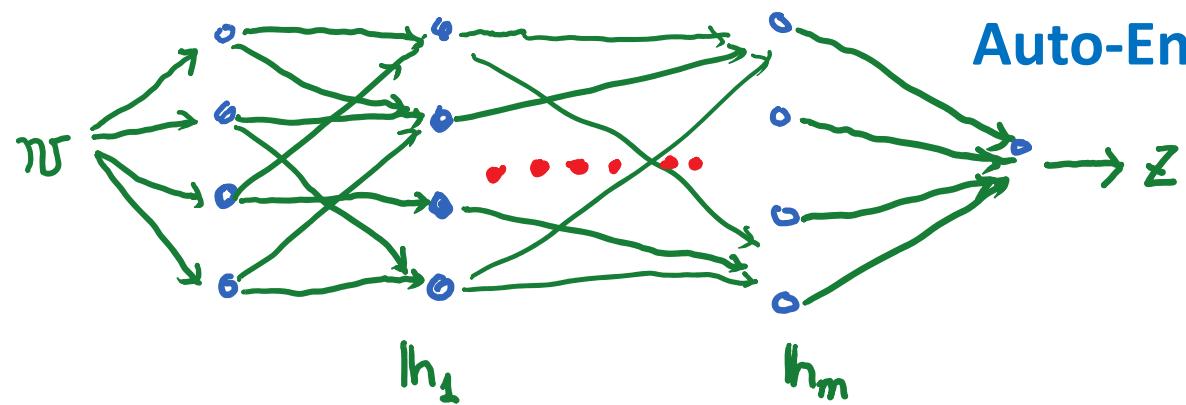
Basic Principles of Unsupervised and Unsupervised Learning Toward Deep Learning

Shun-ichi Amari (RIKEN Brain Science Institute)
collaborators: R. Karakida, M. Okada (U. Tokyo)

Deep Learning

Self-Organization + Supervised Learning

RBM: Restricted Boltzmann Machine
Auto-Encoder, Recurrent Net



tricks !!
ideas !

Dropout
Contrastive divergence

bi-directional .

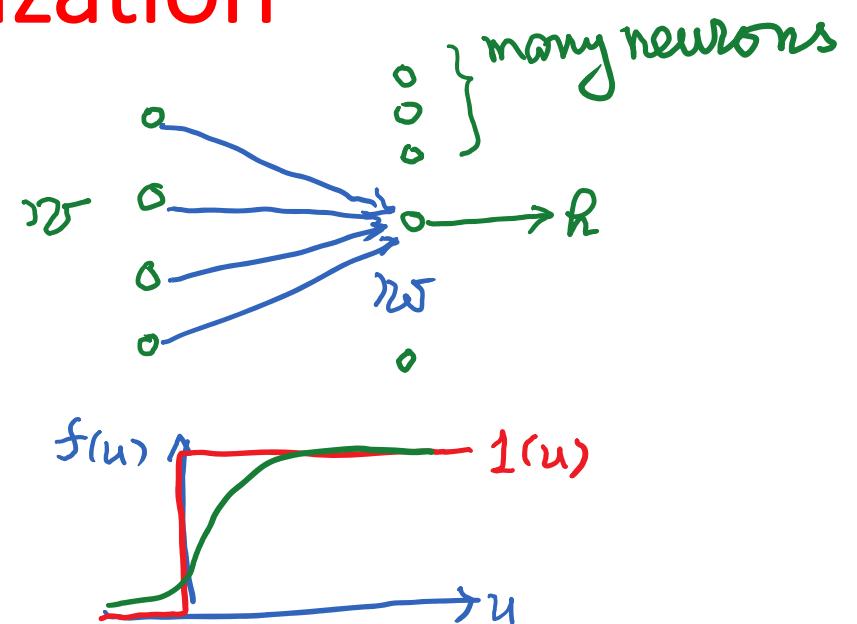
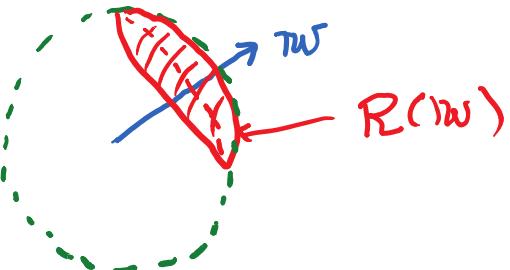
Simple Hebbian Self-Organization

$$h = f(\mathbf{w} \cdot \mathbf{v} - \tau)$$

receptive field

$$R(\mathbf{w}) = \{ \mathbf{v} \mid \mathbf{w} \cdot \mathbf{v} - v_0 > 0 \}$$

$$|\mathbf{w}|^2 = \text{const}$$



self-organization of \mathcal{W} : dynamics of $R(\mathcal{W})$

$$\mathcal{W} : P(\mathcal{W})$$

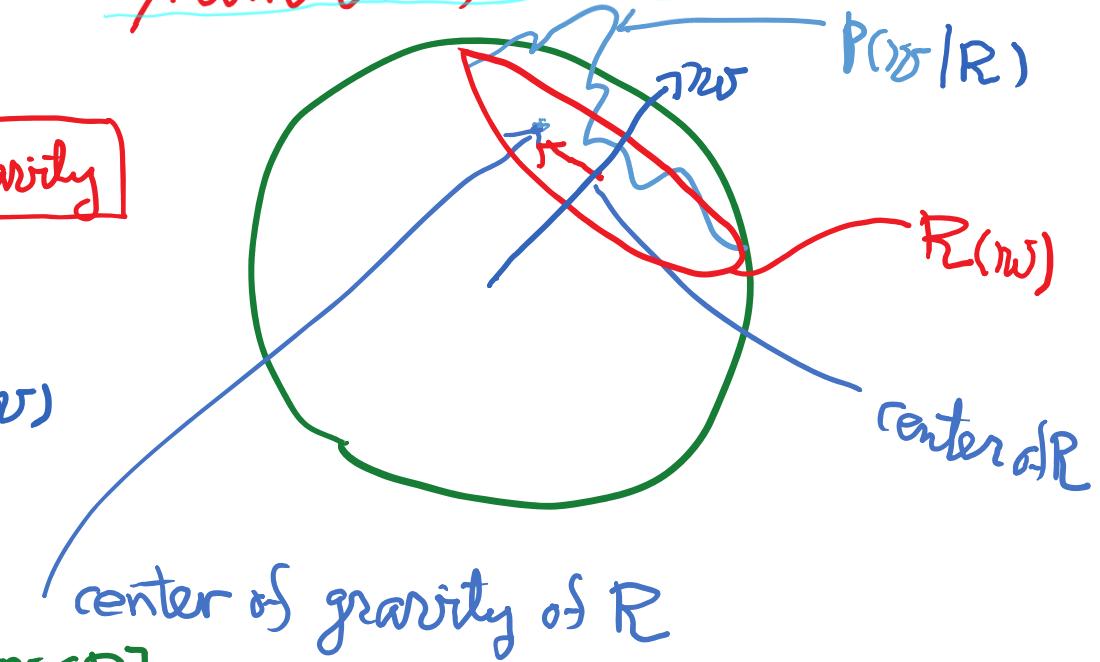
center \Rightarrow C. of gravity

$$\dot{\mathcal{W}} = \langle h(-\mathcal{W} + C\mathcal{W}) \rangle_{P(\mathcal{W})}$$

$$\frac{1}{P_R} \dot{\mathcal{W}} = -\mathcal{W} + C \langle h \mathcal{W} \rangle_{\mathbb{E}}$$

$$\text{Prob}\{\mathcal{W} \in R\} = \langle \mathbb{1}_{\{\mathcal{W} \cdot \mathcal{W} - v_0\}} \rangle_{P(\mathcal{W})}$$

$$E[h\mathcal{W} | \mathcal{W} \in R]$$



$$P(\mathcal{W} | R)$$

$$R(\mathcal{W})$$

center of R

center of gravity of R

Equilibrium

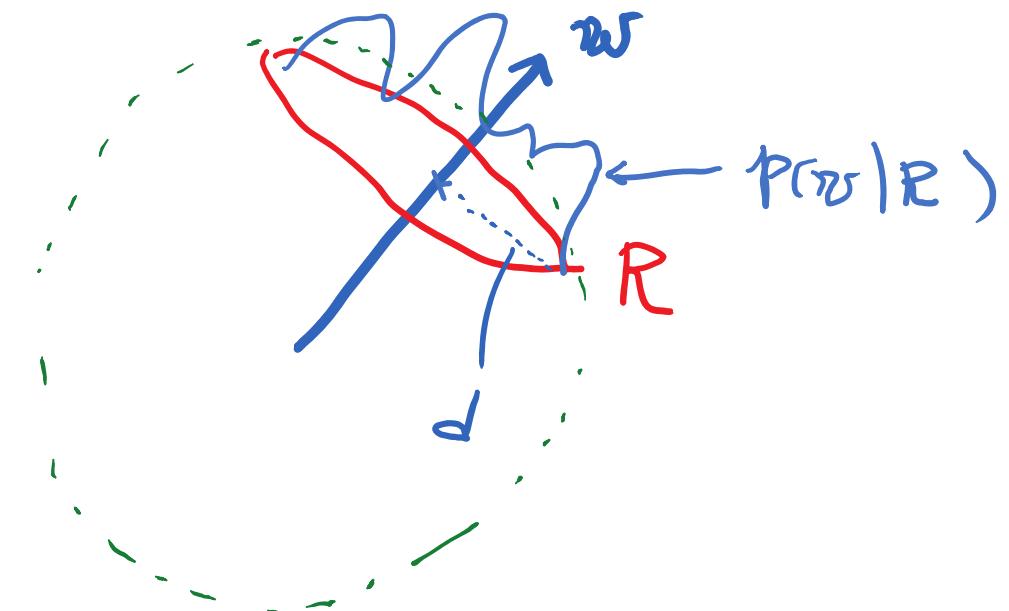
$$\pi\omega = C \langle \rho v \rangle_R$$

\uparrow
center & c. of gravity

$$\pi\omega \cdot \pi\zeta = \zeta$$

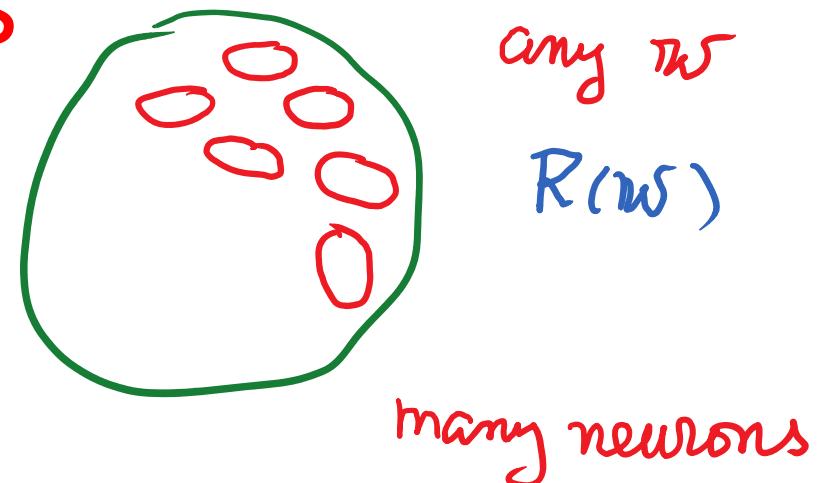
d : radius of $R(\pi\omega)$

ζ : small $\Leftrightarrow d$: large

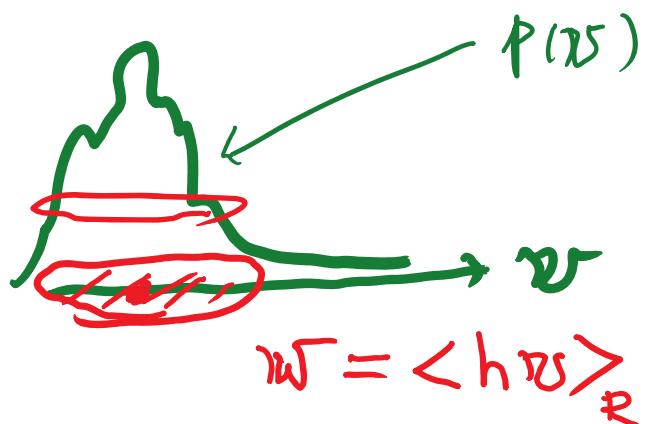


Equilibrium: special cases

1. $P(\eta) = c$: uniform



2. $P(\eta)$: single cluster



Two and many clusters

1. two clusters

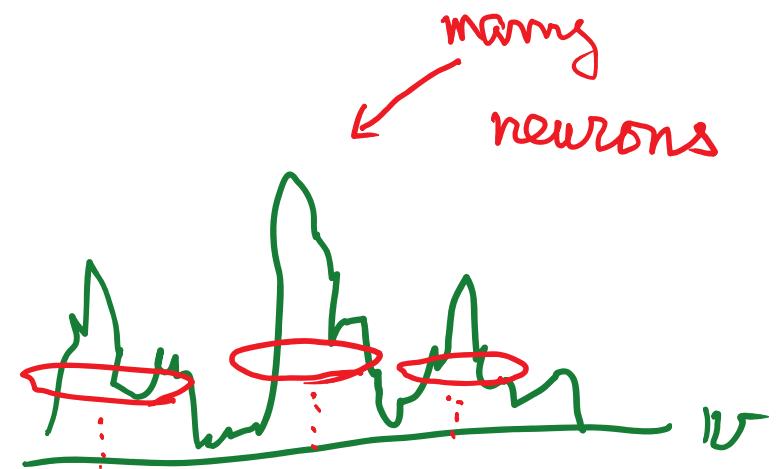
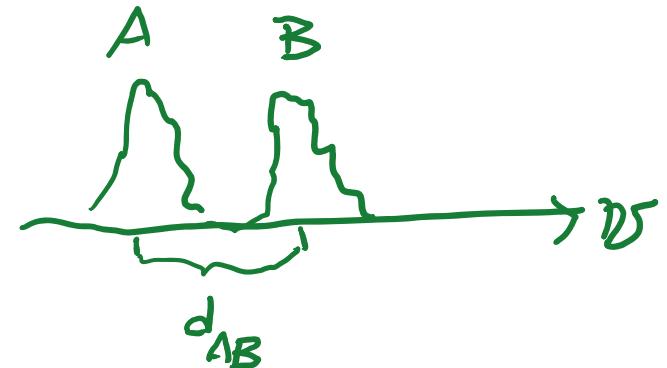
$$d_{AB} > d_0 \quad \mathcal{W} = c \mathcal{V}_A + c \mathcal{V}_B$$

$$d_{AB} < d_0 \quad \mathcal{W} = C_1 \mathcal{V}_A + C_2 \mathcal{V}_B$$

2. many clusters : multi-stable

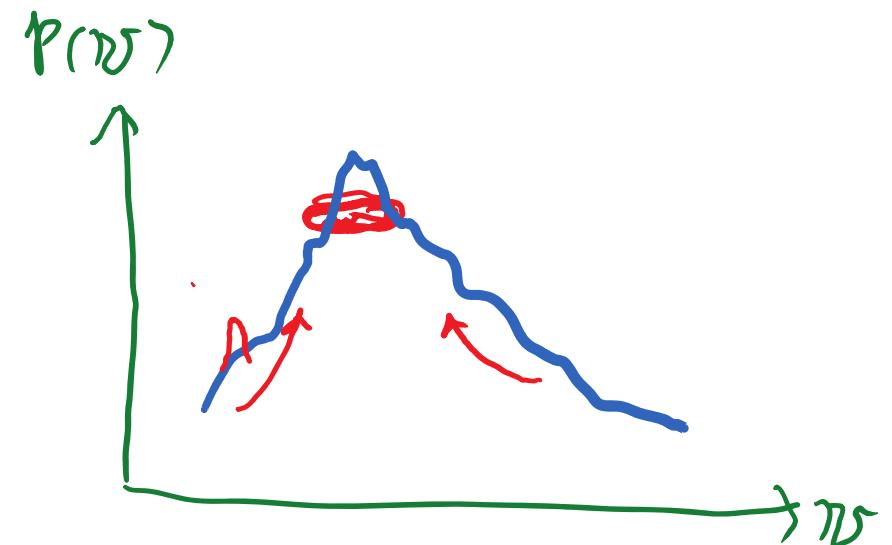
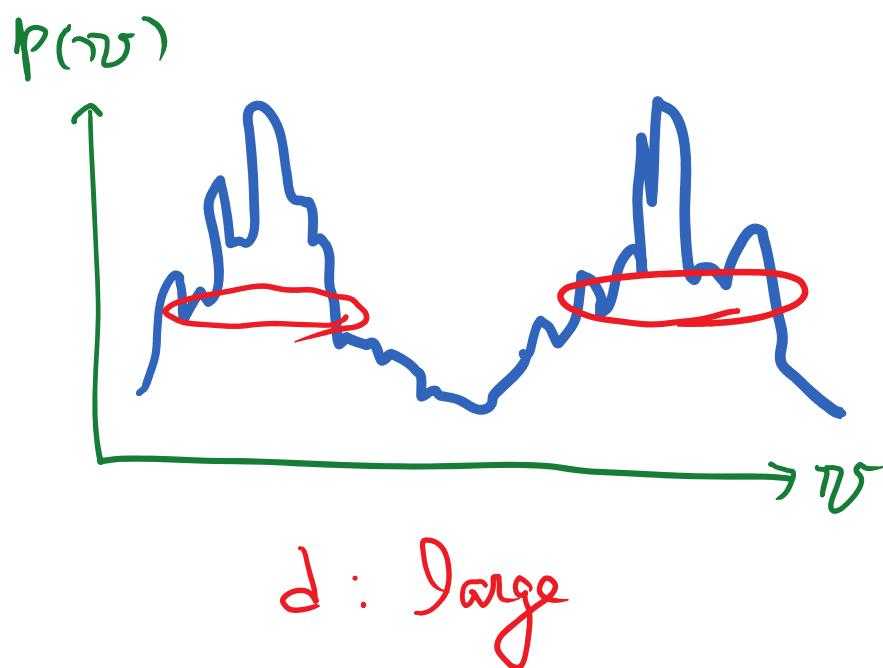
adequate size

center = c. gravity



Dynamics of self-organization

$$\tilde{w} = \langle h w \rangle_k - \bar{w} \approx \nabla p(w) : d \text{ small}$$



Lyapunov Function

$$F(w, \eta) = c(\eta w \cdot v - \tau)^+ - \frac{1}{2} |\eta w|^2$$

$$\dot{w} = h \nabla_w F$$

$$u^+ = \begin{cases} u, & u > 0, \\ 0, & u \leq 0 \end{cases}$$

$$\dot{F} = \nabla F \cdot \dot{w} = h |\nabla F|^2 > 0$$

$$\langle \dot{F} \rangle_{P(w)} = \int |\nabla F|^2 P(w|R) dw > 0$$

convergence to equilibria

Further Problems

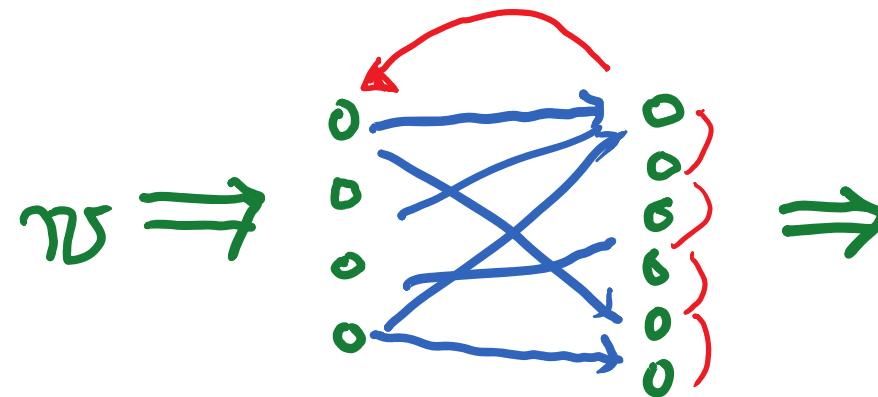
Distributed τ small clusters; large clusters

Mutual interactions among h-neurons neural field

Localized receptive fields

invariance

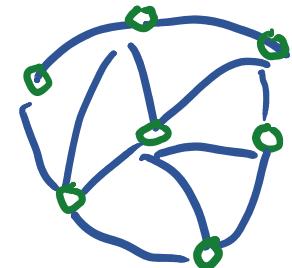
feedback



Boltzmann Machine

Markov chain

$$P(\mathbf{Z}_t \mid \mathbf{Z}_{t-1})$$



$$P(z=1) = f(u), \quad u_i = \sum w_{ij} z_j - b_i$$

$$w_{ij} = w_{ji}$$

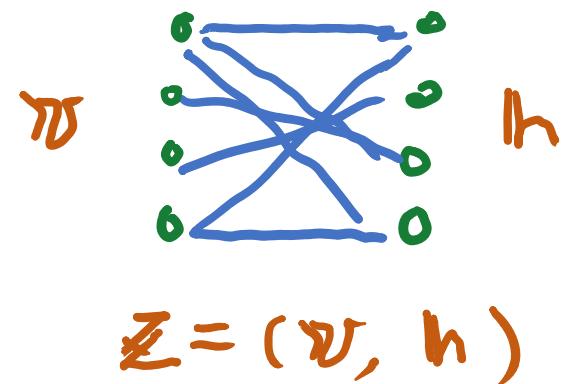
stable distr.

$$P(\mathbf{Z}) = \exp \left\{ \sum s_i z_i + \frac{1}{2} \sum w_{ij} z_i z_j - \psi \right\}$$

RBM: Restricted Boltzmann Machine

$$P(\mathbf{v}, \mathbf{h}) = \exp \left\{ b \cdot \mathbf{v} + c \cdot \mathbf{h} + \mathbf{h}^\top W \mathbf{v} - \psi \left(-\frac{1}{2} \|\mathbf{v}\|^2 - \frac{1}{2} \|\mathbf{h}\|^2 \right) \right\}$$

\uparrow \uparrow
 $\sum h_i w_{ij} v_j$ $\sum v_i^2$

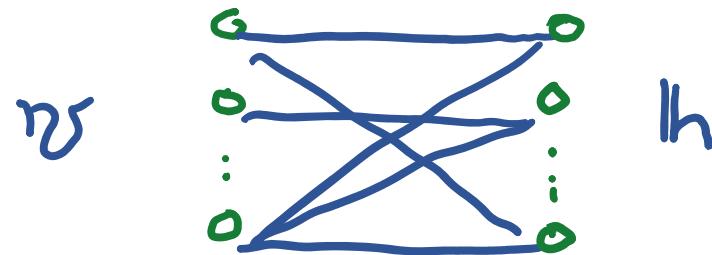


energy machine :

b, c : neglect

RBM

$$g(v, h) = \exp \{ h^T W v - \psi \}$$



marginal distribution : $\tilde{g}(v) = \sum_h g(v, h) \approx p(v)$

$$\tilde{g}_k(h) = \sum_v g(v, h)$$

conditional distribution

$$g(h|v) = \prod_i f(\sum_j w_{ij} v_j - \tau_i),$$

conditionally
independent

$$g(v|h) = \prod_j f(\sum_i h_i w_{ij} - \tau_j),$$

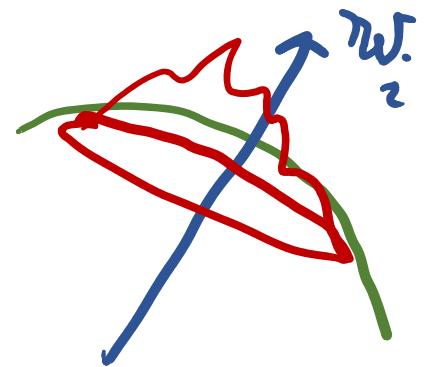
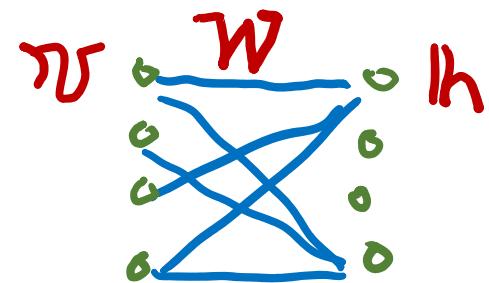
Self-Organization

minimize $_{w} \text{KL} [p(v) : g_v(v; w)]$

$$= \int p(v) \log \frac{p(v)}{g_v(v; w)} dv$$

$$dw_i = \varepsilon \left\{ - \underbrace{\langle h_i v \rangle_g}_{\text{center } w_i ?} + \underbrace{\langle h_i v \rangle_p}_{\text{center of gravity of } P(w_i)} \right\}$$

$$\langle h v \rangle_g = \frac{\partial}{\partial v} \Psi(v)$$

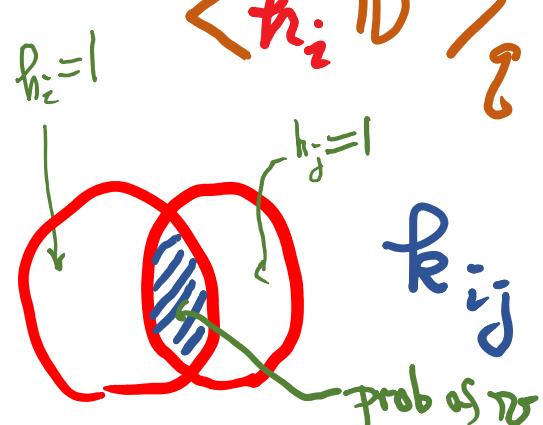


Interaction of Hidden Neurons

$$\langle h_i v \rangle_g = \frac{\partial \Psi(W)}{\partial w} : g(h, v) = \exp\{h^T W v - \Psi(W)\}$$

$$\begin{aligned}\Psi(W) &= \log \sum_m \left(\exp\left\{-\frac{1}{2} \|Wv\|^2 + h^T W v\right\} dv \right) \\ &= \log \sum_m \exp\left\{\frac{1}{2} \|h^T W\|^2\right\} + C\end{aligned}$$

$$\langle h_i v \rangle_g = \sum R_{ij} w_j : \text{interaction}$$



$$R_{ij} = E_g[h_i h_j] \leftarrow \text{joint firing prob.}$$

v : analog, Gaussian

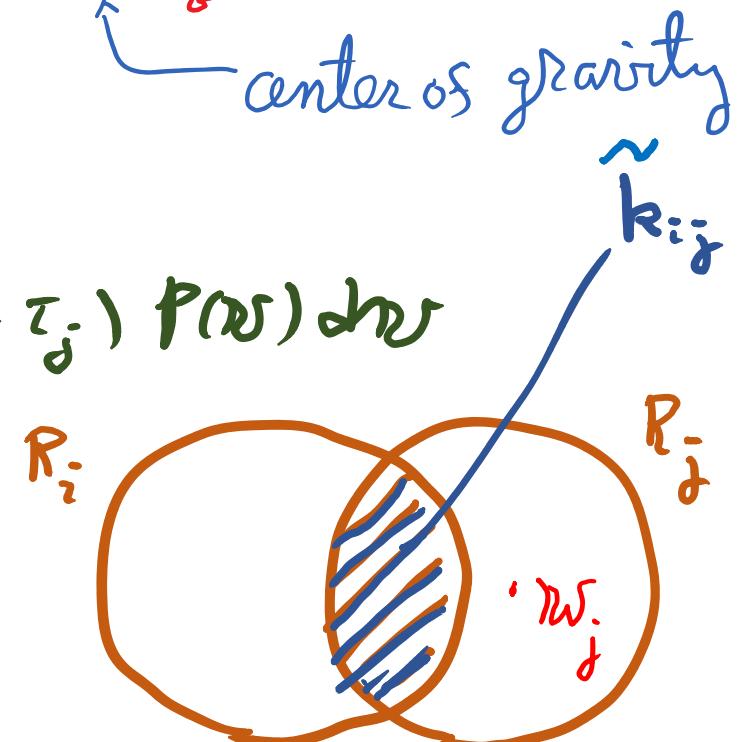
$$w_j = \sum (\hat{R}^{-1})_{ji} \langle h_i v \rangle_{\text{cg}}$$

v : Gaussian

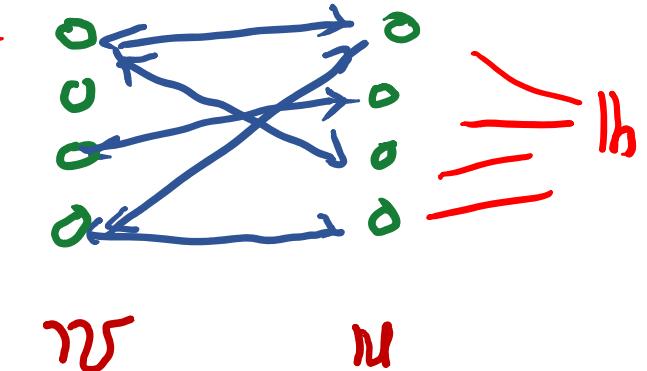
$$k_{ij} = \int f(v_i \cdot v - \tau_i) f(v_j \cdot v - \tau_j) P(v) dv$$

$$\langle h_i v \rangle_{\text{cg}} = \sum \tilde{k}_{ij} w_j$$

↑
center gravity
interaction



Recurrent Net (Auto-Encoder)

$$\begin{cases} \dot{u} = -u + Wv \\ \dot{v} = -\nu + S^T h + a \end{cases} \quad \left. \begin{array}{l} \text{multi stable} \\ p(a) \dots a_b \end{array} \right.$$


$$\dot{w} = -w + c \langle lh v^T \rangle_{p(a)}$$

$$\dot{s} = -s + c \langle dv h^T \rangle_{p(a)}$$

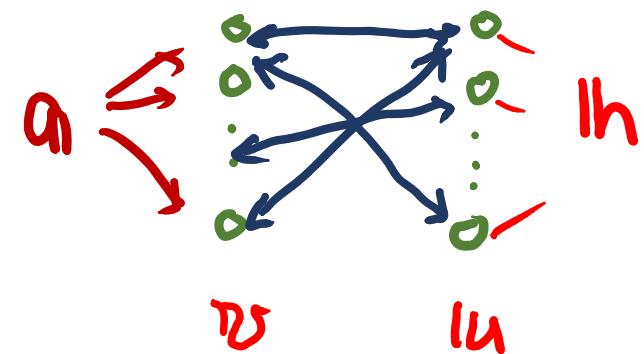
$$s_{ij} = w_{ij} \quad (\text{equilibrium})$$

Recurrent Net Self-Organization

$$w_i = c \left(\langle h_i | a \rangle_{\text{pr}(a)} + c' \sum_j k_{ij} w_j \right)$$

↑ center of gravity ↑ interaction

$$k_{ij} = \langle h_i | h_j \rangle_{\text{pr}(a)}$$



Gaussian RBM is easy

$$\log p(v, h) = h^T W v - \psi + \log \frac{p(v)}{g(v)}$$

$$\log g(v, h) = h^T W v - \psi$$

Higher-order interactions

Gram-Charlier expansion

$$\langle h | v \rangle_{r, g}$$

Gaussian world

Gaussian Boltzmann Machine

$$g(\mathbf{v}, \mathbf{h}) = \exp\left\{-\frac{1}{2}|\mathbf{v}|^2 - \frac{1}{2}|\mathbf{h}|^2 + \mathbf{h}^\top \mathbf{W} \mathbf{v} - \Psi(\mathbf{w})\right\}$$

$$\langle \mathbf{h} \mathbf{v} \rangle_p = \mathbf{w} \mathbf{c}$$

center of gravity

$$\mathbf{c} = \int \mathbf{v} \mathbf{v}^\top p(\mathbf{v}) d\mathbf{v}$$

covariance matrix

$$\langle |\mathbf{h} \mathbf{v}| \rangle_i = (\mathbf{I} - \mathbf{W} \mathbf{W}^\top)^{-1} \mathbf{W}$$

interactions of h. neurons

Equilibrium Solution

$$WC = (I - WW^T)^{-1} W$$

$$W = \begin{pmatrix} w_1 \\ \vdots \\ w_k \end{pmatrix}$$

solution : $w_i \cdot w_j = 0 \quad (i \neq j), \quad \|w_i\|^2 = m_i$

$$w_i C = \underbrace{(I - m_i)}_{\lambda_i}^{-1} w_i$$

$$w_i C = \lambda_i w_i$$

PCA!

$$m_i = 1 - \frac{1}{\lambda_i}$$

Equilibrium Solution

$$WC = (I - WW^T)^{-1}W$$

General Solution

$$W = U \text{diag} \left(\sqrt{1 - \frac{1}{\lambda_1}}, \dots, \sqrt{1 - \frac{1}{\lambda_m}}, 0, \dots, 0 \right) V$$

- orthogonal matrix : U, V
- C diagonalized by V
$$C = V^T \text{diag}(\lambda_1, \dots, \lambda_n) V$$

You can choose $m (\leq k)$ eigen values form $\lambda_1, \dots, \lambda_k > 1$

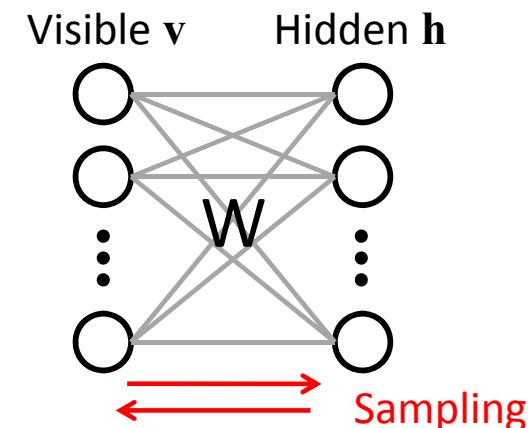
Stable Solution the case of $m = k$

Contrastive Divergence

- RBM

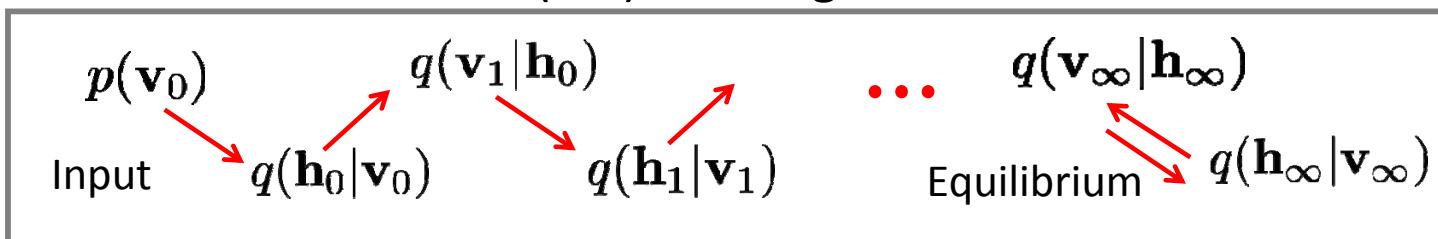
- 2-layered probabilistic neural network
- No connections within layers

$$q(\mathbf{h}, \mathbf{v}; W) = \exp(-\mathbf{h}^T W \mathbf{v}) / \sum_{\mathbf{h}, \mathbf{v}} \exp(-\mathbf{h}^T W \mathbf{v})$$



- How to train RBM

Maximum Likelihood (ML) learning is hard



Many iterations of Gibbs Sampling demand **too much computational time**

Contrastive Divergence Solution

$$\langle h \nabla \rangle_{\tilde{q}} = W C W^T W + W$$

$$W C = W C W^T W + W$$

$$w_i \cdot C = \frac{1}{1 - m_i} w_i$$

PCA !

CD_n Solution

$$WC = W(W^T W)^n C (W^T W)^n + W \sum_{k=0}^{2n-1} (W^T W)^k$$

General Solution

$$W = U \text{diag} \left(\sqrt{1 - \frac{1}{\lambda_1}}, \dots, \sqrt{1 - \frac{1}{\lambda_m}}, 0, \dots, 0 \right) V$$

Stable Solution

the case of $m = k$

the same analytical form with maximum likelihood
regardless of “n”

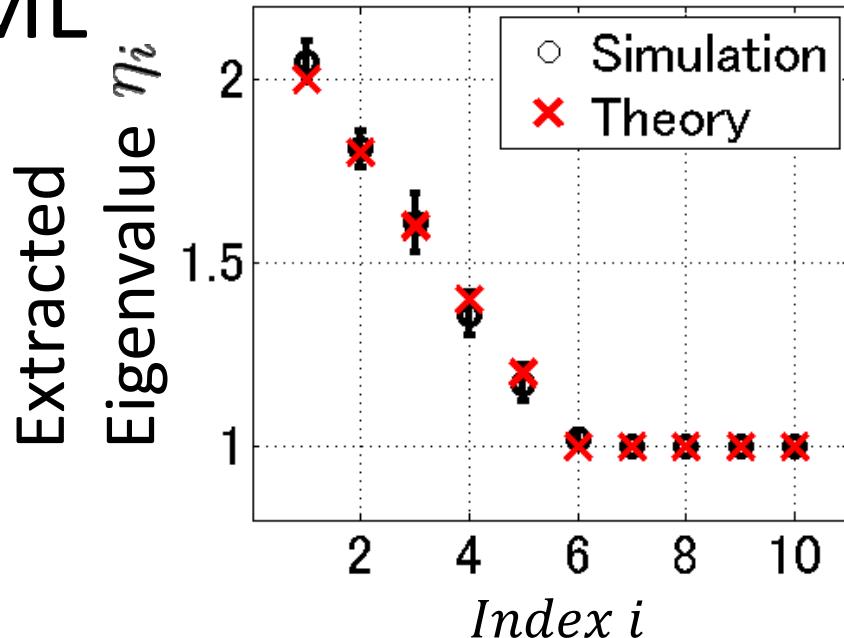
Simulation

Each Layer : 10 Neurons

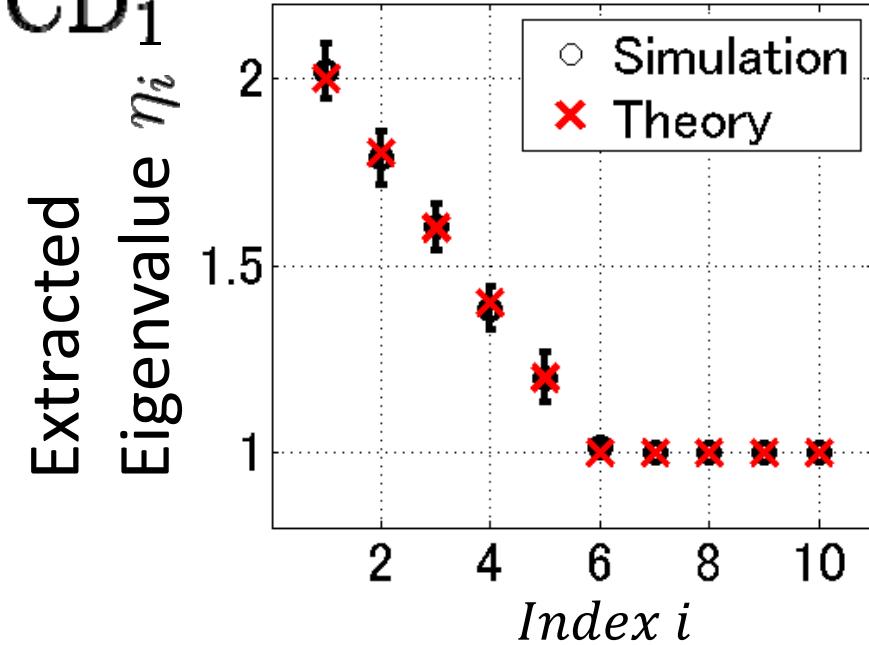
Input: 10-dim. Gaussian Distribution

Mean = 0, Variance[0.2, 0.4, ..., 2], Covariance = 0

ML



CD₁



Extracted

Input Eigenvalues

$$\lambda_i = (0.2, 0.4, \dots, 2)$$



Extracted Eigenvalues

$$\eta_i = (1, \dots, 1, 1.2, \dots, 2)$$

Bayesian Duality in Exponential Family

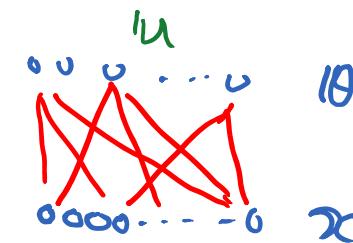
Data x Parameter (higher-order concepts) θ

$$P(x|\theta) = \exp\{\theta \cdot x - \bar{h}(x) - \bar{\psi}(\theta)\}$$

$$p(x, \theta) = \exp\{\theta \cdot x - \bar{h}(x) - \bar{\psi}(\theta)\} \quad \psi - \log \pi(\theta)$$

$$p(\theta|x) = \exp\{\theta \cdot x - \bar{h}(x) - \bar{\psi}(\theta)\}$$

Curved exponential family



$$\mathcal{X}(v), \quad \Theta(u) : \exp\{\theta(u) \cdot \mathcal{X}(v) - \bar{h} - \bar{\psi}\}$$

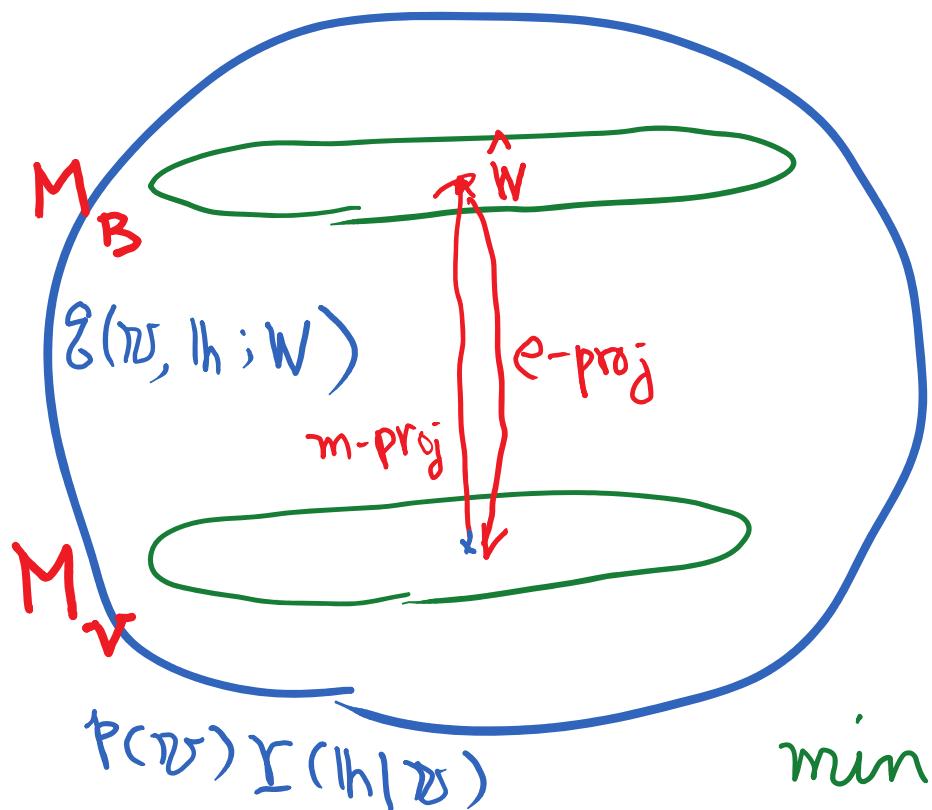
RBM : curved exp. family

$$\Theta = h, \quad x = Wv$$

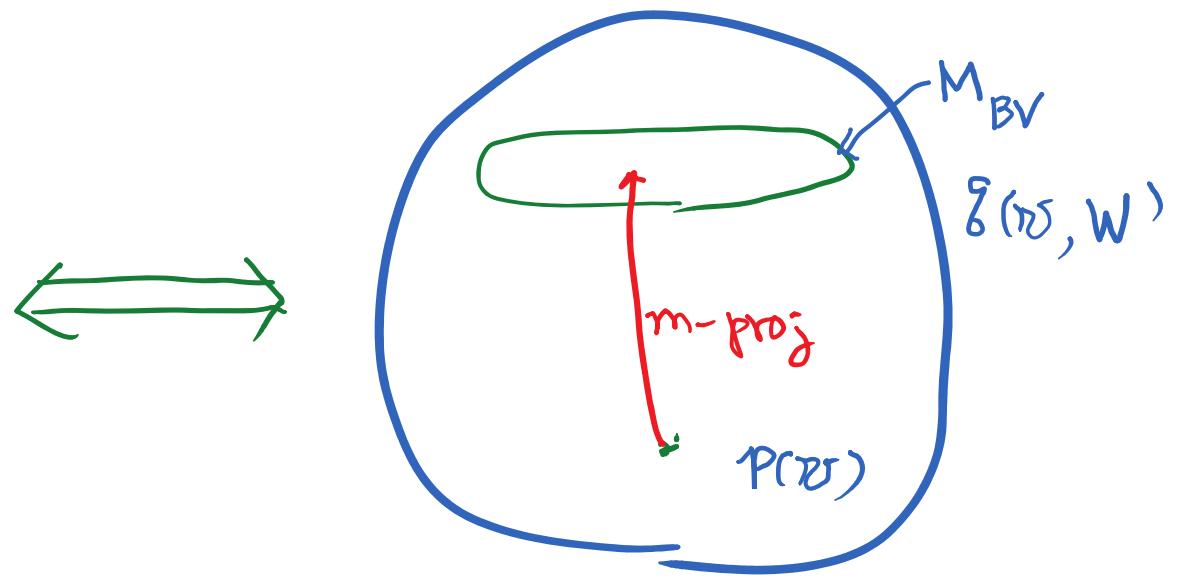
$$x = v \quad \Theta = hW$$

$$P(\mathcal{D}, lh) = \exp \{ \Theta \cdot x - \bar{k} - \bar{\Phi} \}$$
$$lh^T W \mathcal{D}$$

Two Manifolds

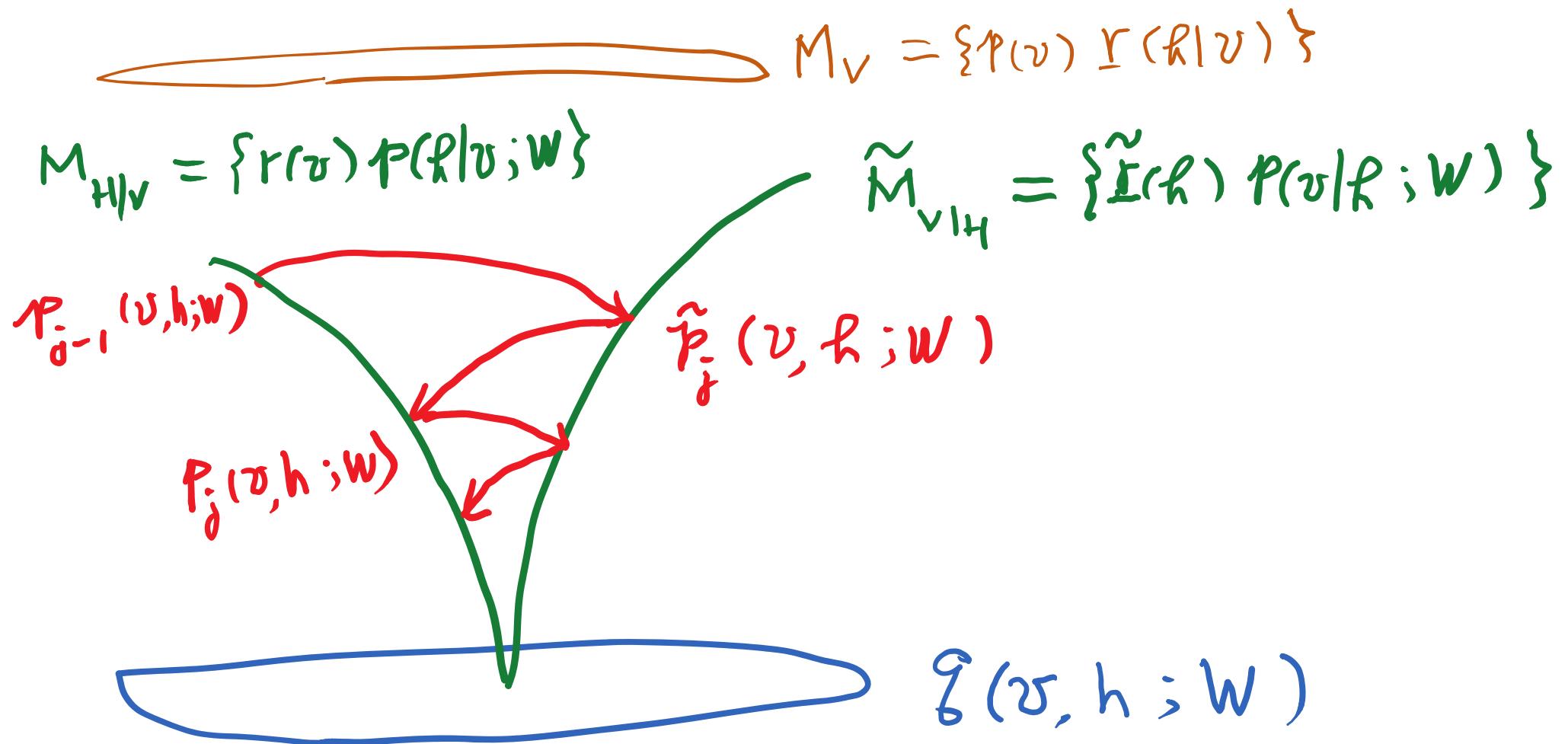


$$\min_w KL[p(v) : g_v(v; w)]$$



$$\min_{\Sigma, W} KL[p(v) \Gamma(h|v) : g(v, h; W)]$$

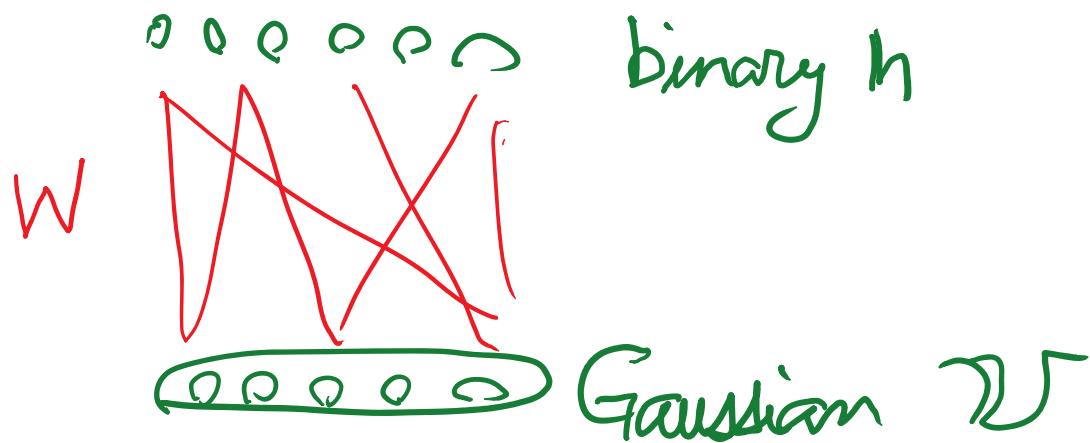
Geometry of CDn (contrastive divergence)



Bernoulli-Gaussian RBM

ICA

R. Karakida



$$p(\mathbf{v}) : \mathbf{v} = \underline{\text{ODS}}$$

$$\mathbf{W} = \mathbf{A}^{-1}$$

Equilibrium Analysis: Results

Assumption of Input $p_0(\mathbf{v})$

$$\underline{\mathbf{v} = B\mathbf{s}} \quad p(\mathbf{s}) = p(s_1)p(s_2) \cdots p(s_N) \quad s_i \geq 0$$

s : Independent and nonnegative sources

B: $N \times N$ orthogonal matrix

ICA (independent Component Analysis) Solutions

If $|\mu_i| \gg \sigma$, ML and CD_n learning have the following stable solutions:

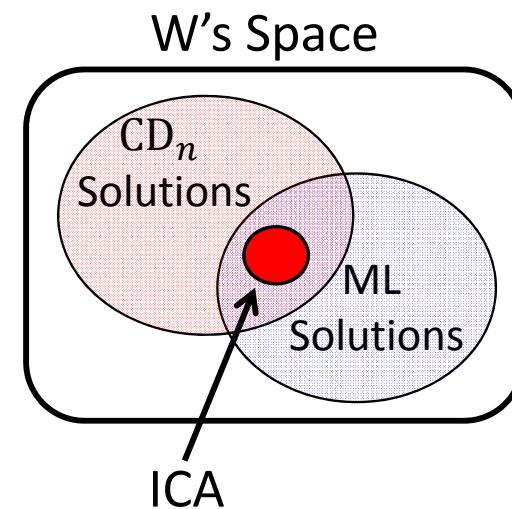
$$\bar{W} = DB^T$$

$$D = \sigma^{-1} \text{diag}(\mu_1, \mu_2, \dots, \mu_N)$$

$$\text{Mean value: } \mu_i = \int s_i p(s_i) ds_i$$

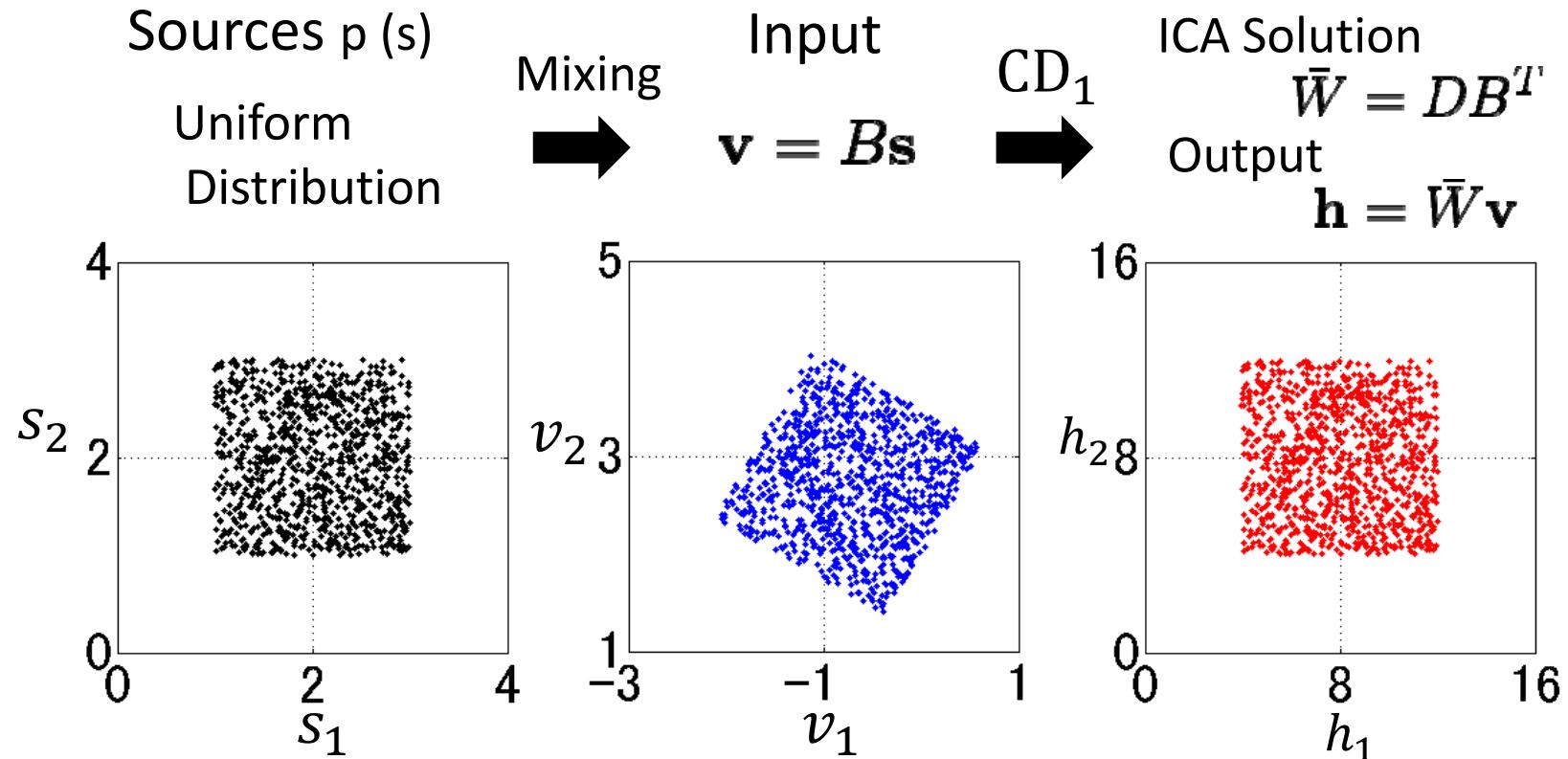
$$\text{Model variance : } \sigma$$

$$\mathbf{h} = \bar{W}\mathbf{v} = DB^T B\mathbf{s} = D\mathbf{s} \quad h_i \propto s_i$$



Simulation

The number of Neurons: $N = M = 2$, $\sigma = 1/2$



Independent sources are extracted in G-B RBM

Supervised Learning

Multilayer perceptron

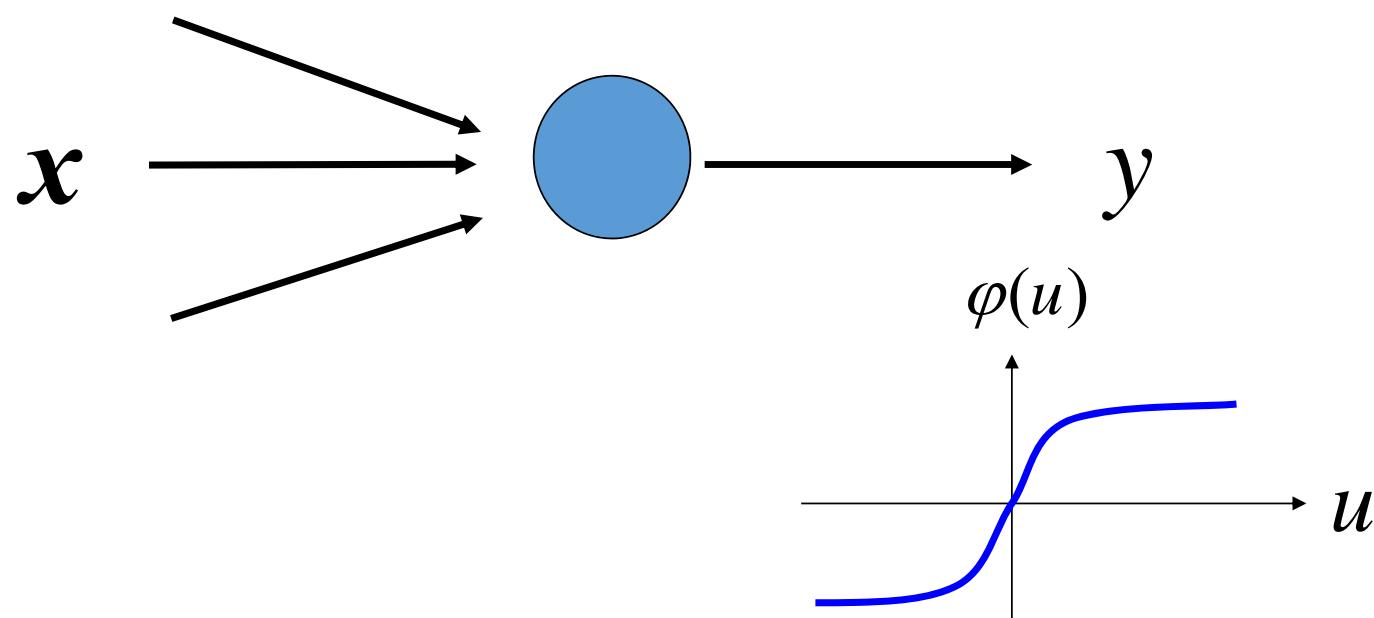
Back-prop learning

Singularity!!

Natural Gradient Solves Difficulty

Mathematical Neurons

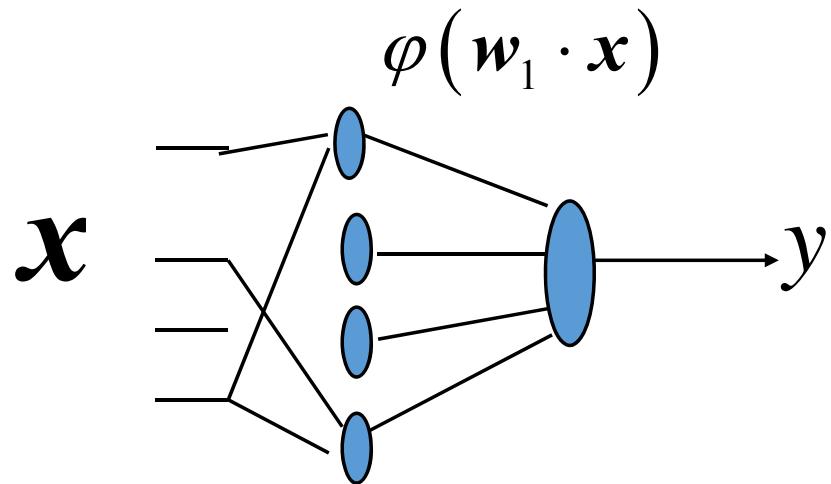
$$y = \varphi\left(\sum w_i x_i - h\right) = \varphi(w \cdot x)$$



Multilayer Perceptrons

$$y = \sum v_i \varphi(w_i \cdot x)$$

$$x = (x_1, x_2, \dots, x_n)$$



$$f(x, \theta) = \sum v_i \varphi(w_i \cdot x)$$

$$\theta = (w_1, \dots, w_m; v_1, \dots, v_m)$$

Multilayer Perceptron

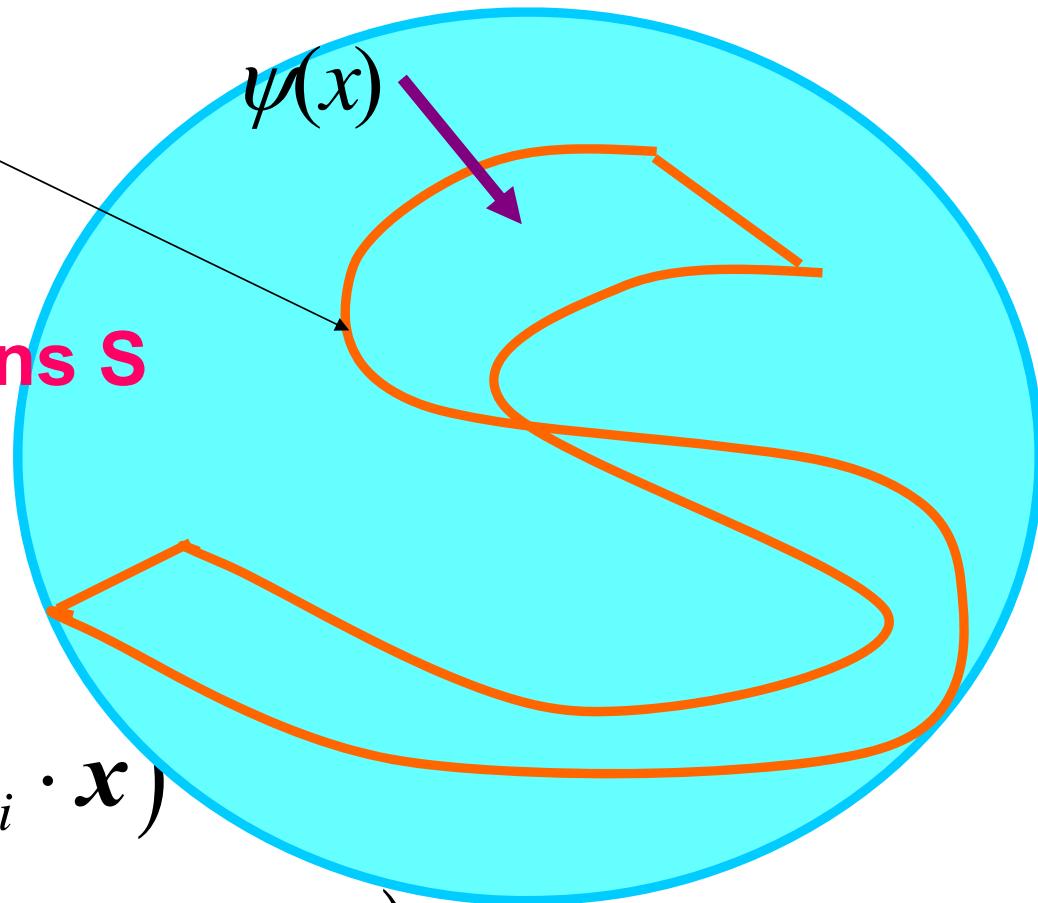
neuromaniold

space of functions S

$$y = f(x, \theta)$$

$$= \sum v_i \varphi(w_i \cdot x)$$

$$\theta = (v_1, \dots, v_m ; w_1, \dots, w_m)$$



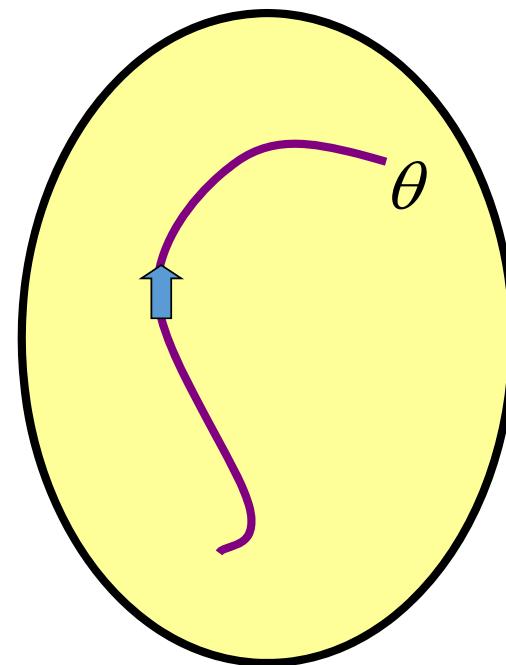
Backpropagation ---gradient learning

examples: $(y_1, \mathbf{x}_1), \dots (y_t, \mathbf{x}_t)$ -- training set

$$l(y, x; \theta) = \frac{1}{2} |y - f(\mathbf{x}, \theta)|^2$$
$$= -\log p(y, \mathbf{x}; \theta)$$

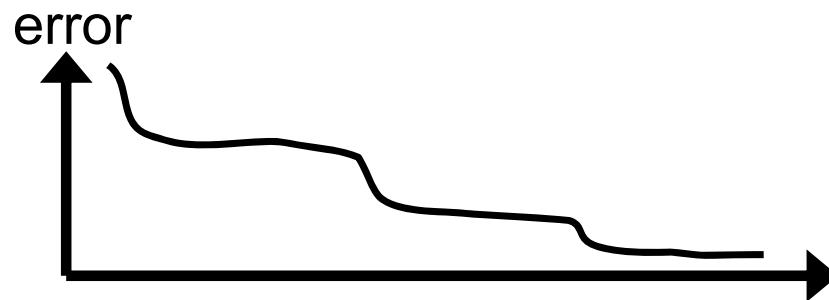
$$\Delta \theta_t = -\eta_t \frac{\partial l(y_t, \mathbf{x}_t; \theta_t)}{\partial \theta}$$

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



Flaws of MLP

slow convergence : Plateau

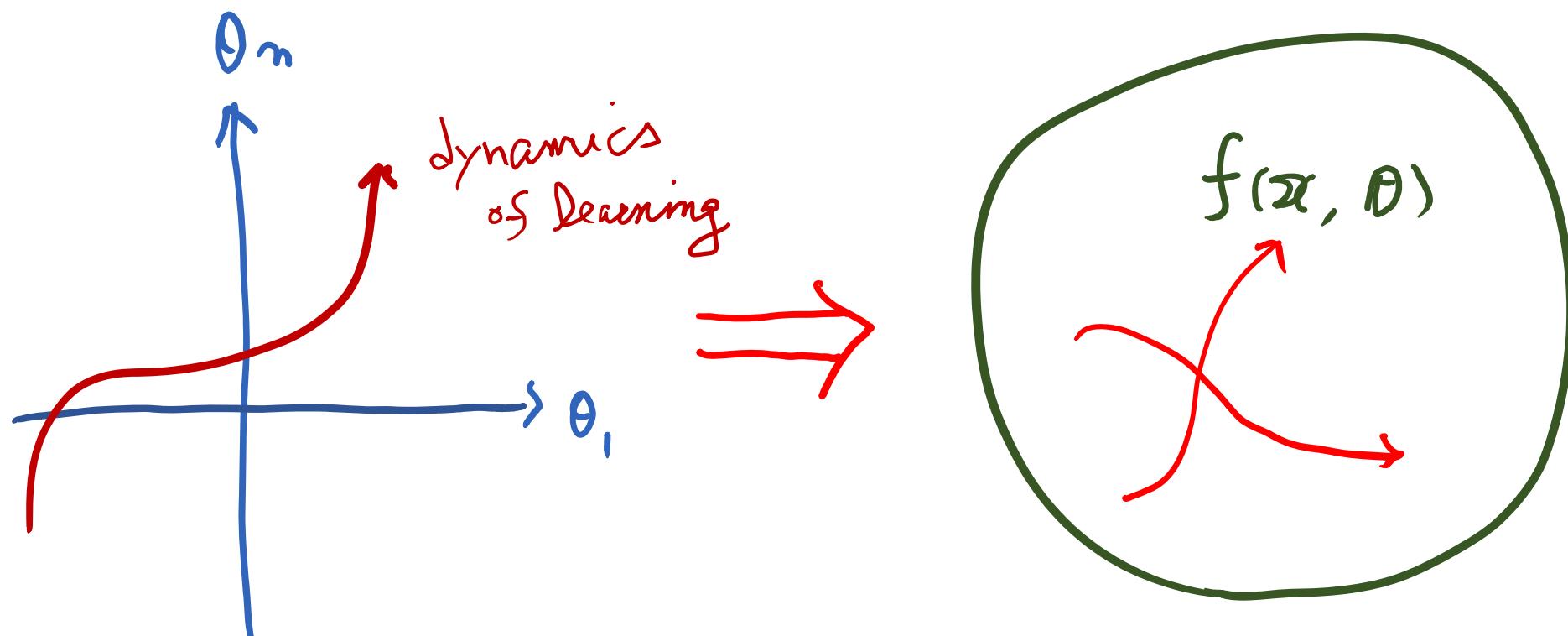


local minima



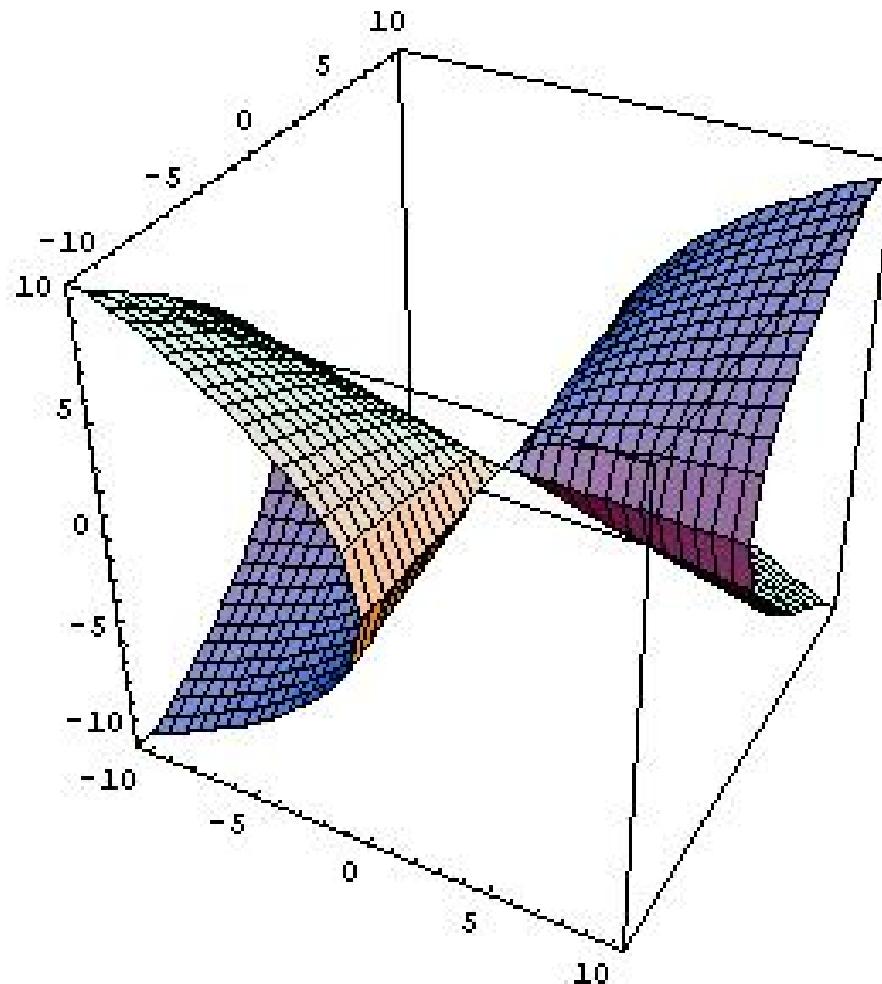
→ Boosting and Bagging; SVM

Parameter Space vs Function Space



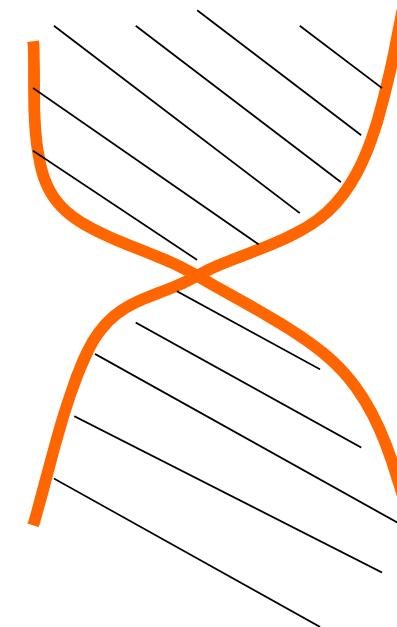
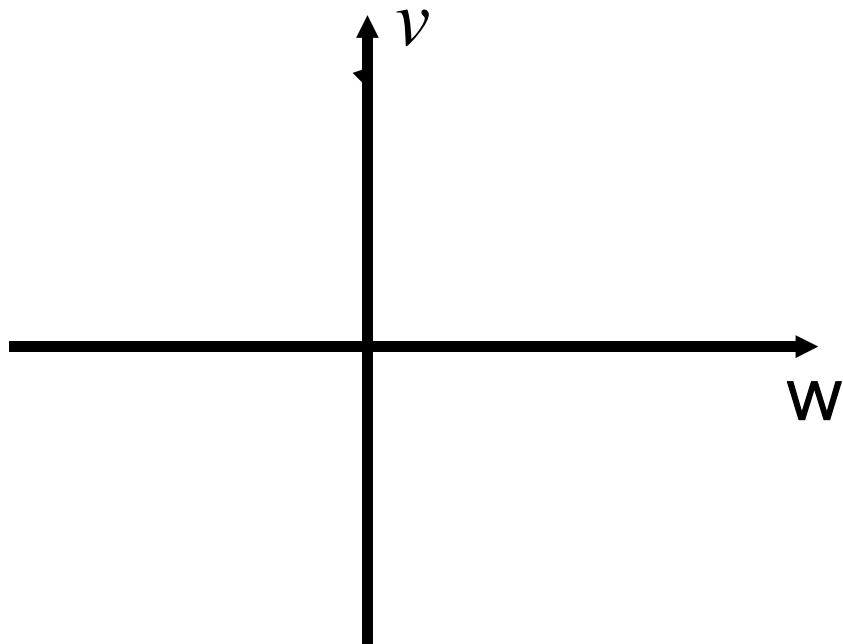
$$\theta = (\theta_1, \dots, \theta_n)$$

Singularity of MLP--example



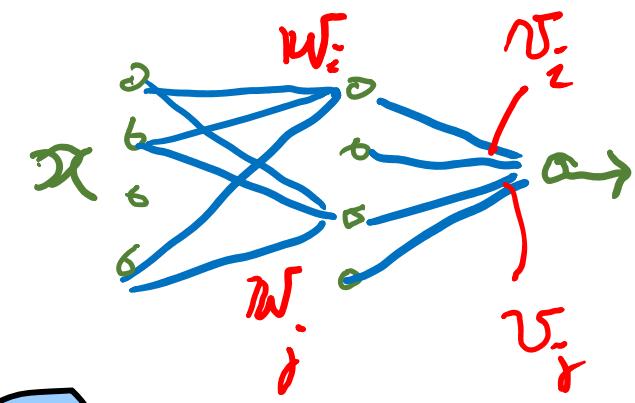
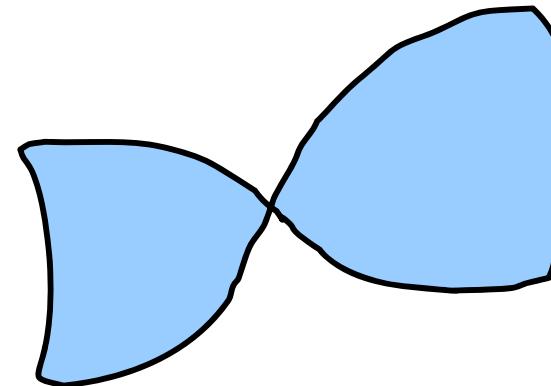
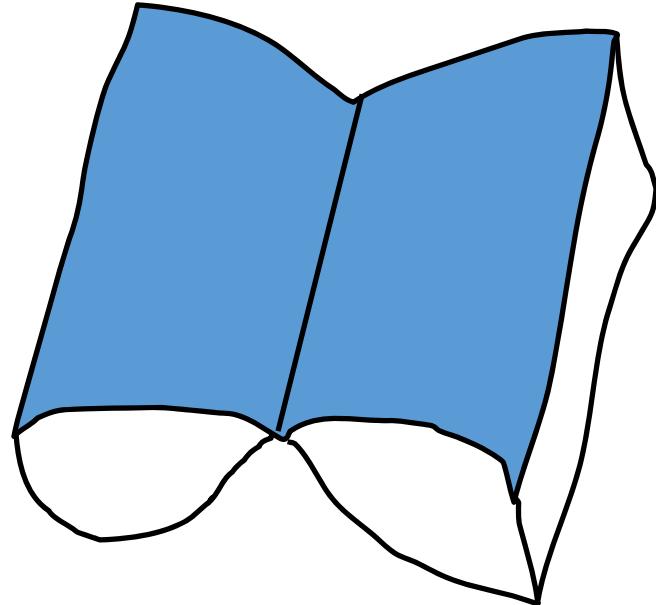
Geometry of singular model

$$y = v\varphi(w \cdot x) + n \quad v | w | = 0$$



singularities

$$\begin{cases} v_i = 0 \\ w_i = w_j \end{cases}$$

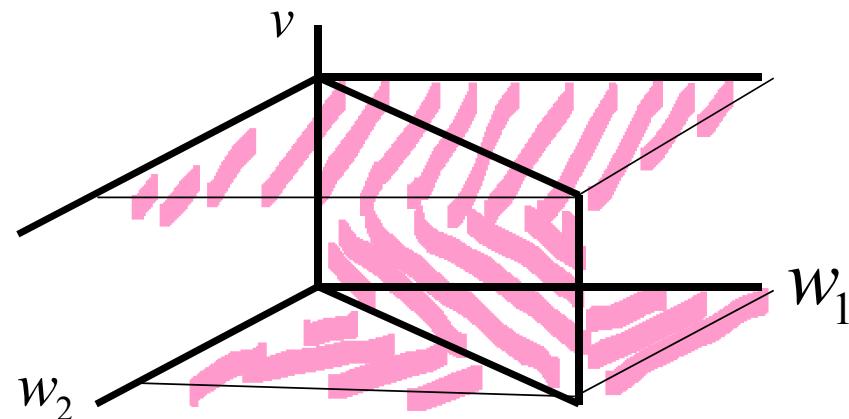


Gaussian mixture

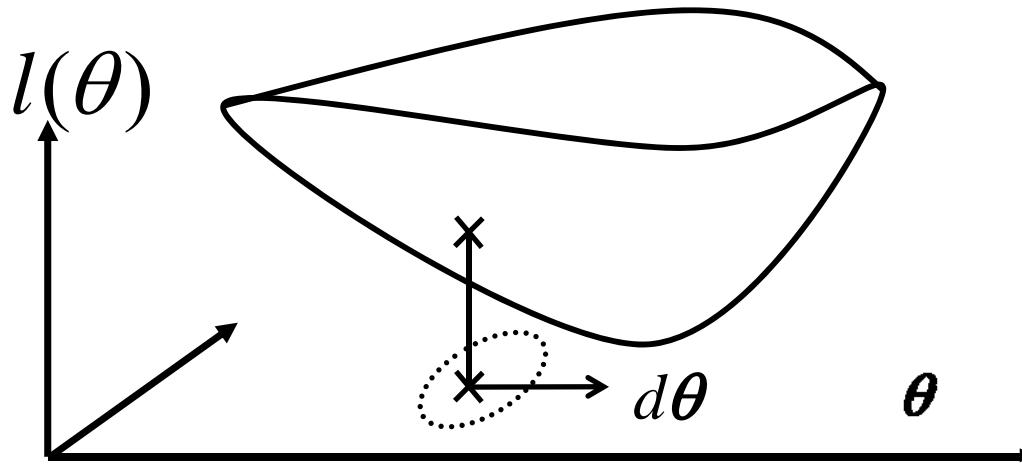
$$p(x; \nu, w_1, w_2) = (1 - \nu)\varphi(x - w_1) + \nu\varphi(x - w_2)$$

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

singular: $w_1 = w_2, \quad \nu(1 - \nu) = 0$



Steepest Direction---Natural Gradient



$$\nabla l = \left(\frac{\partial l}{\partial \theta_1}, \dots, \frac{\partial l}{\partial \theta_n} \right) \quad \Delta \theta_t = -\eta_t \nabla l(x_t, y_t; \theta_t)$$

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

$$|d\theta|^2 = d\theta^T G d\theta = \sum G_{ij} d\theta^i d\theta^j$$

Natural Gradient

$$\max \quad d l = l(\theta + d\theta) - l(\theta)$$

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

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

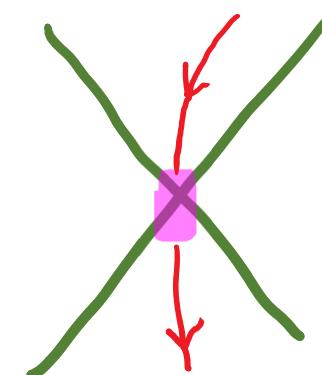
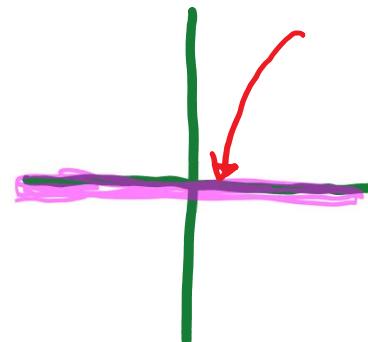
$$\Delta \theta_t = -\eta_t \tilde{\nabla} l(x_t, y_t; \theta_t)$$

Adaptive Natural Gradient

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

$G^{-1} \rightarrow \infty$, $\nabla l \rightarrow 0$ at singularities

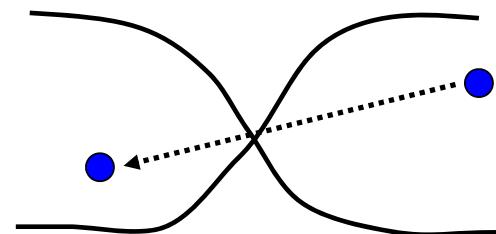
$G^{-1} \nabla l$



Learning, Estimation, and Model Selection

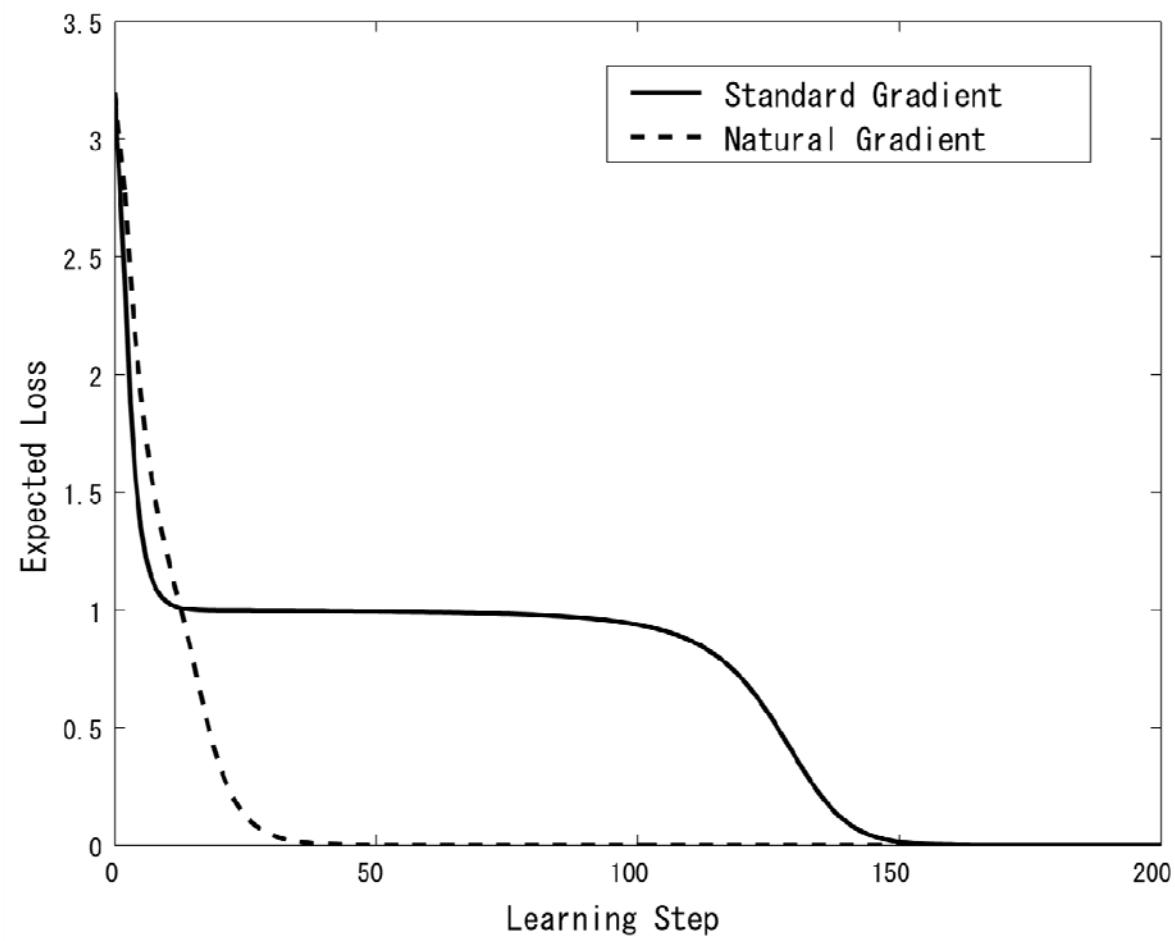
$$E_{\text{gen}} = D \left[p_0(y|x) : p(y|x; \hat{\theta}) \right]$$

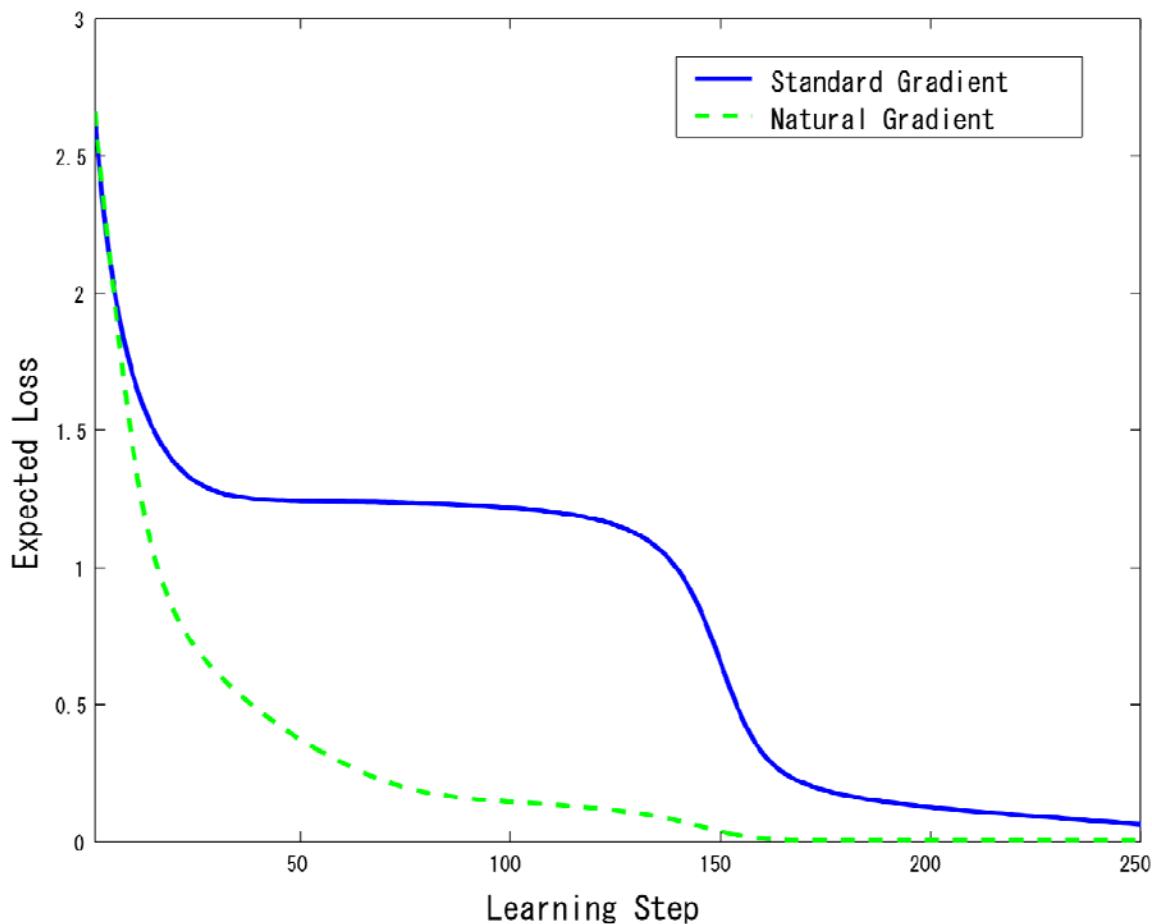
$$E_{\text{train}} = D \left[p_{\text{emp}}(y|x; \hat{\theta}) \right]$$

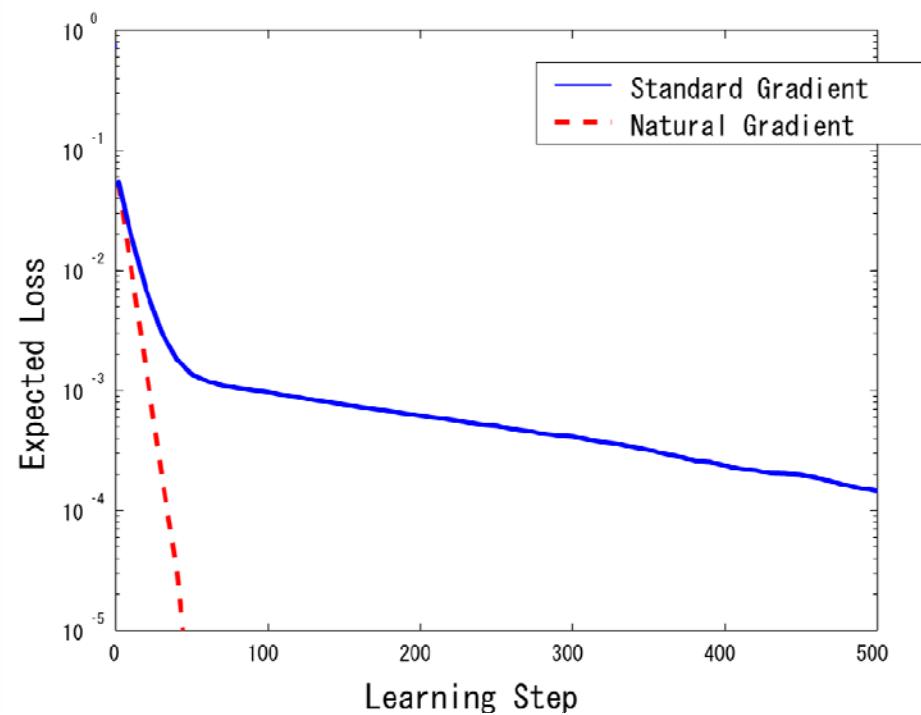


$$E_{\text{gen}} = \frac{d}{2n} \quad d : \text{dimension}$$

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

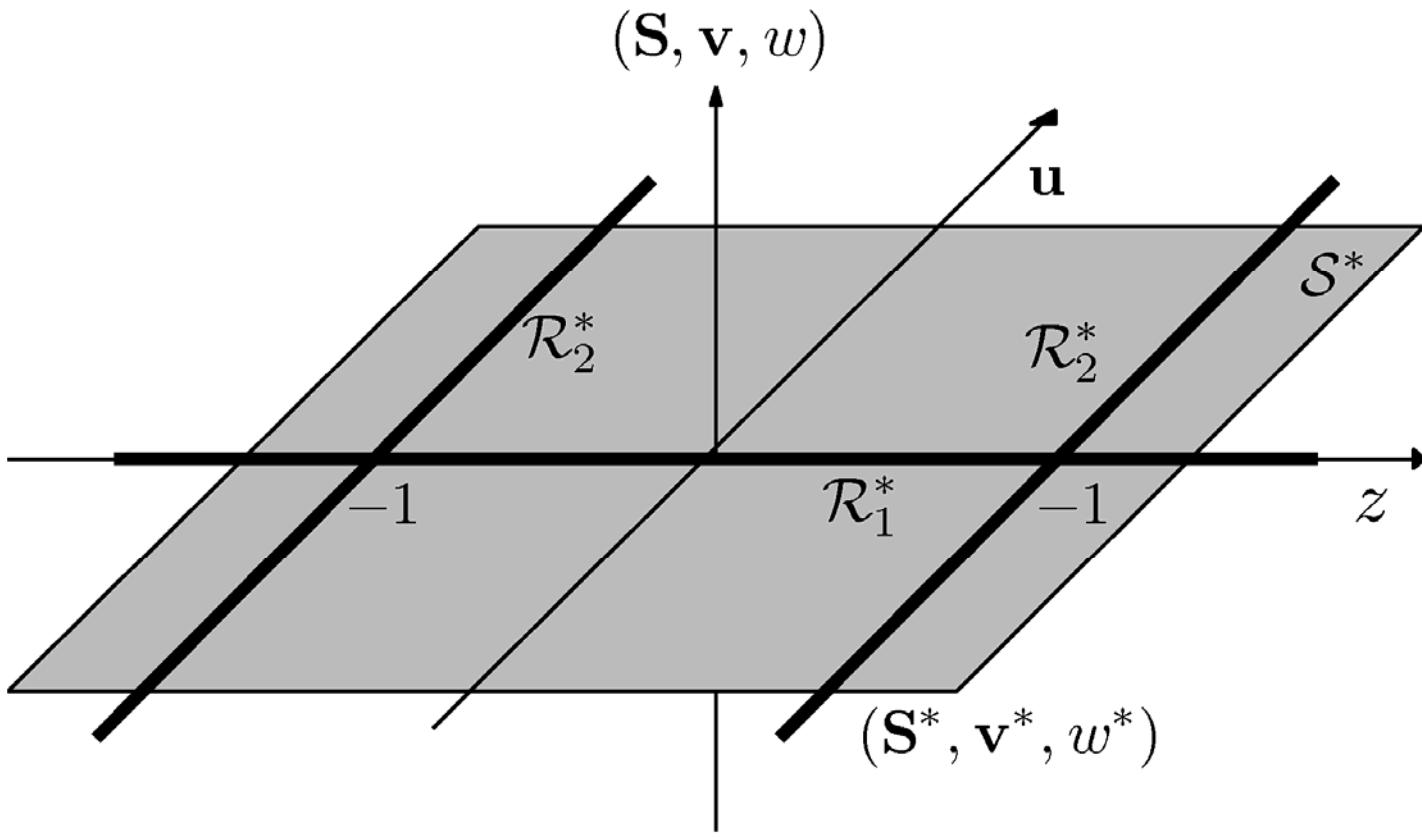






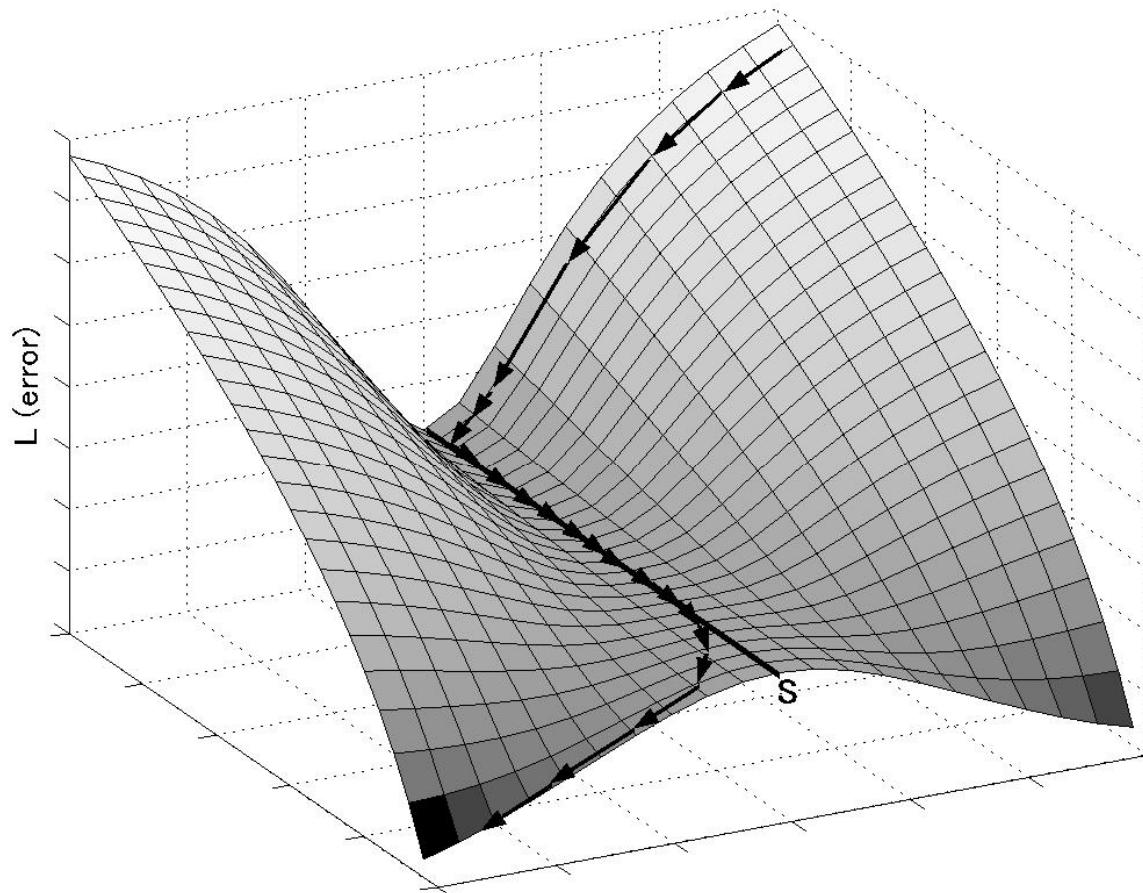
Coordinate Transformation

$$\left\{ \begin{array}{ll} \mathbf{u} = \mathbf{w}_2 - \mathbf{w}_1 & : \mathbf{u} = 0 \\ \mathbf{w} = \frac{\nu_1 \mathbf{w}_1 + \nu_2 \mathbf{w}_2}{\nu} & \mathbf{w} = \mathbf{w}^* \\ \nu = \nu_1 + \nu_2 & \nu = \nu^* \\ z = \frac{\nu_2 - \nu_1}{\nu} & z = \pm 1 \end{array} \right. \quad \mathcal{R}_1 \quad \mathcal{R}_2$$



Singular lines in the parameter space

Learning Trajectory near the singularity



Milnor attractor

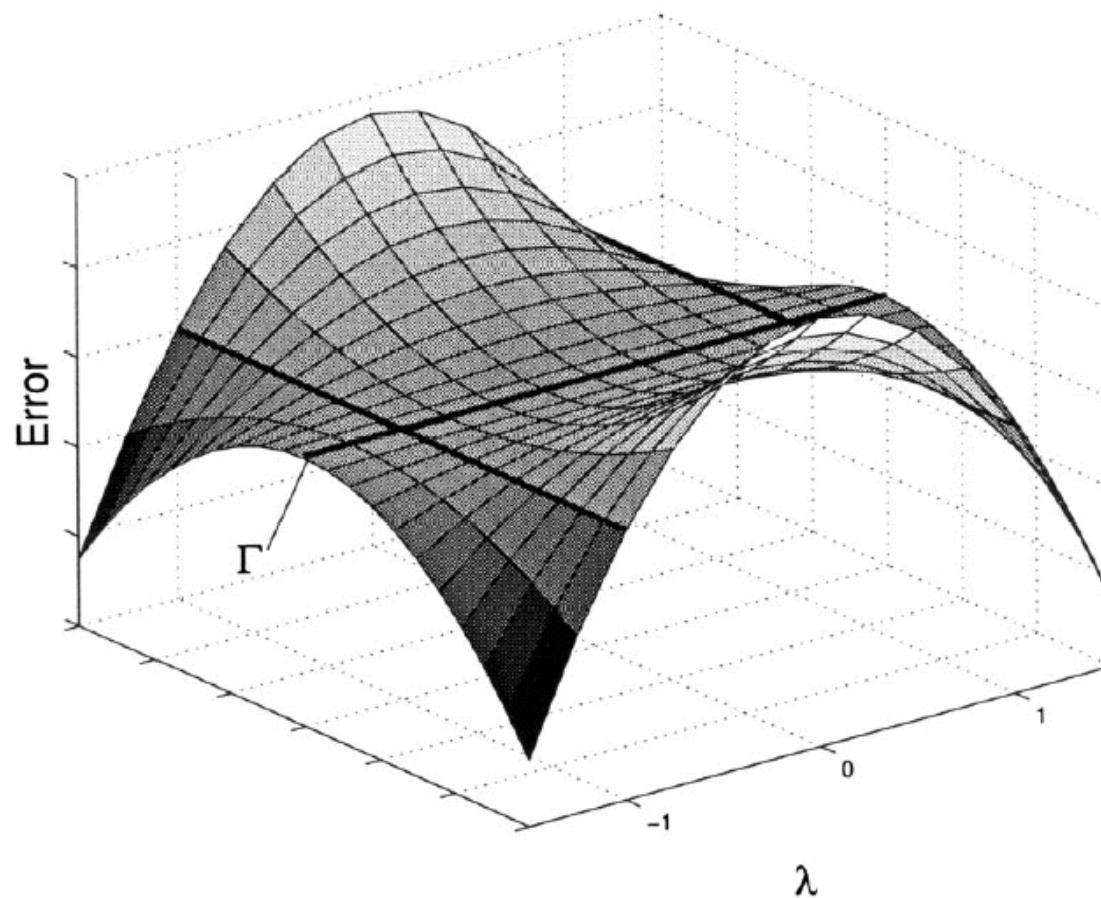
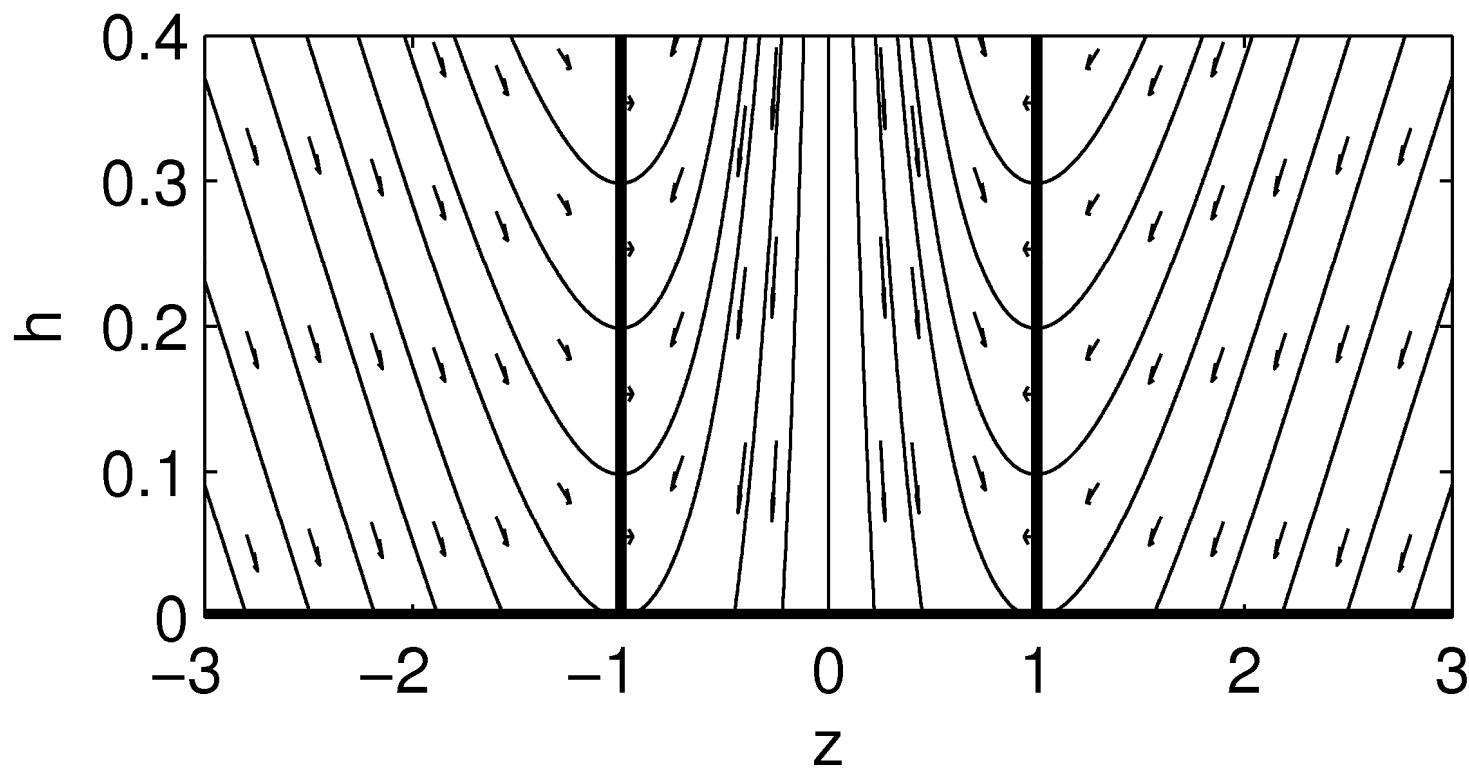


Fig. 5. Critical set with local minima and plateaus.



Dynamic vector fields: Redundant case

Dynamics of Learning : Trajectories

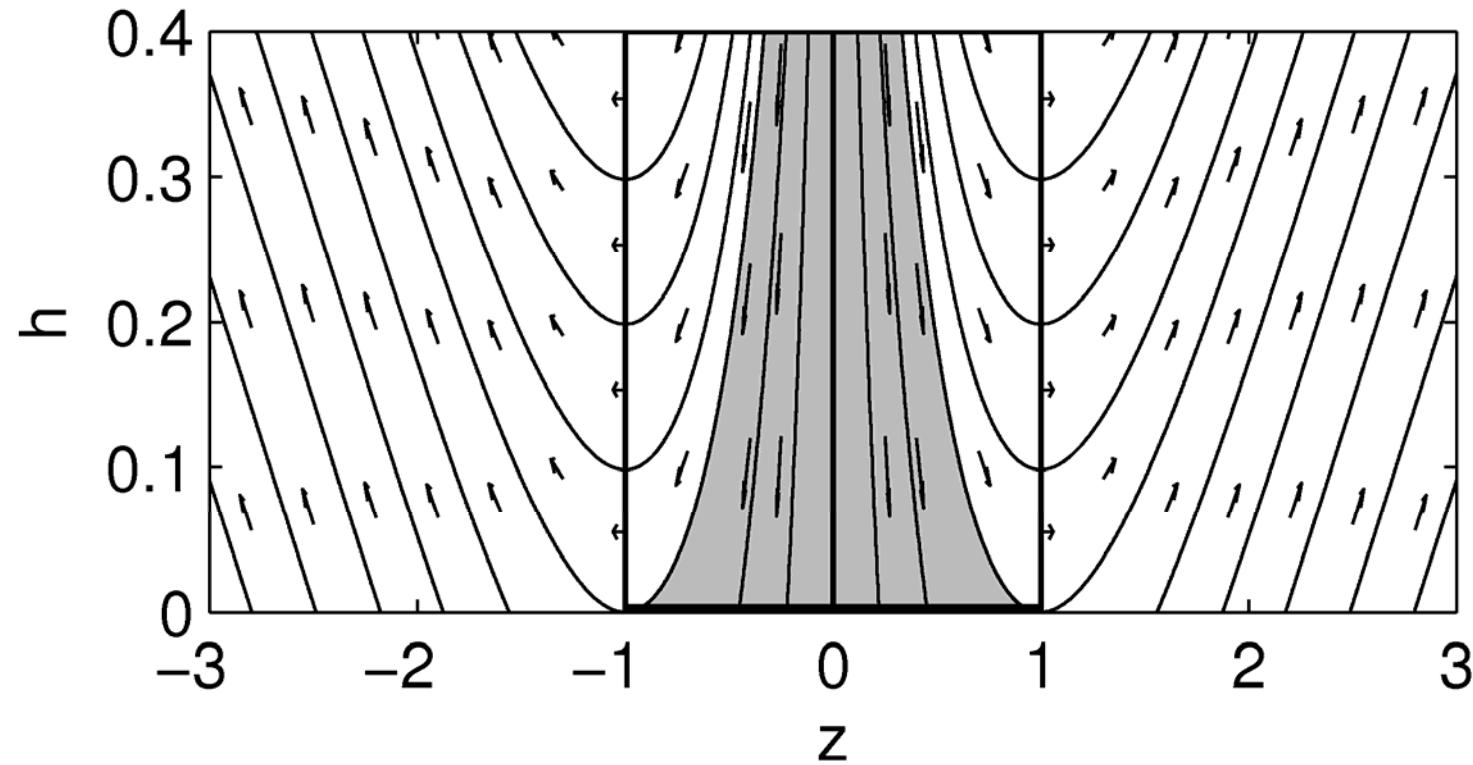
$$\dot{\boldsymbol{u}} = 2\nabla_{\boldsymbol{u}} l$$

$$\dot{z} = \frac{2(z^2 + 1)}{v^{*2}} \nabla_z l$$

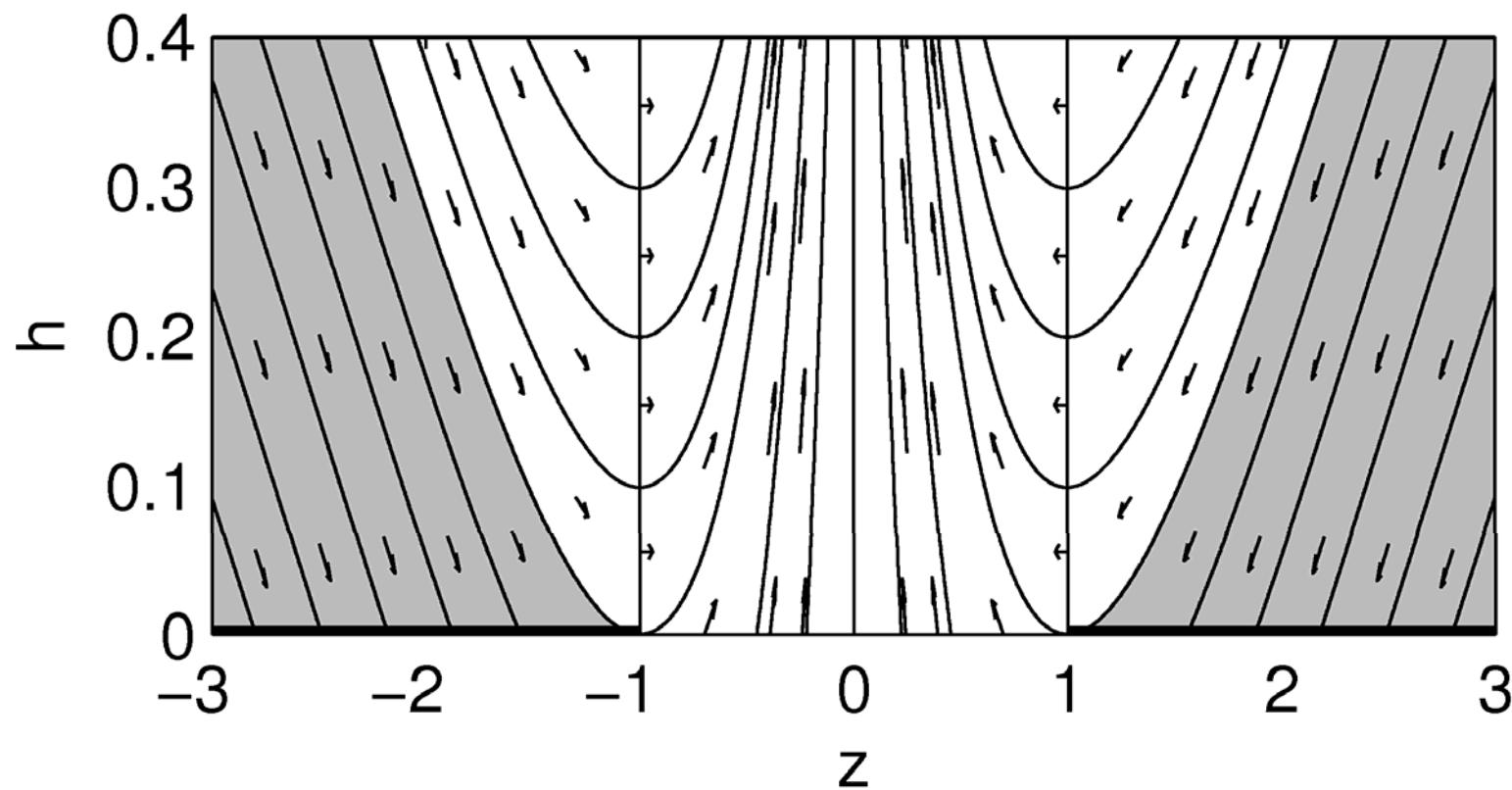
$$h(\boldsymbol{u}) = \frac{1}{2} |\boldsymbol{u}|^2$$

$$\dot{h} = \frac{v^{*2} (z^2 - 1)}{z(z^2 + 1)} z$$

$$h(\boldsymbol{u}) = v^{*2} \log\left(|z| + \frac{1}{|z|}\right) + c$$



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



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

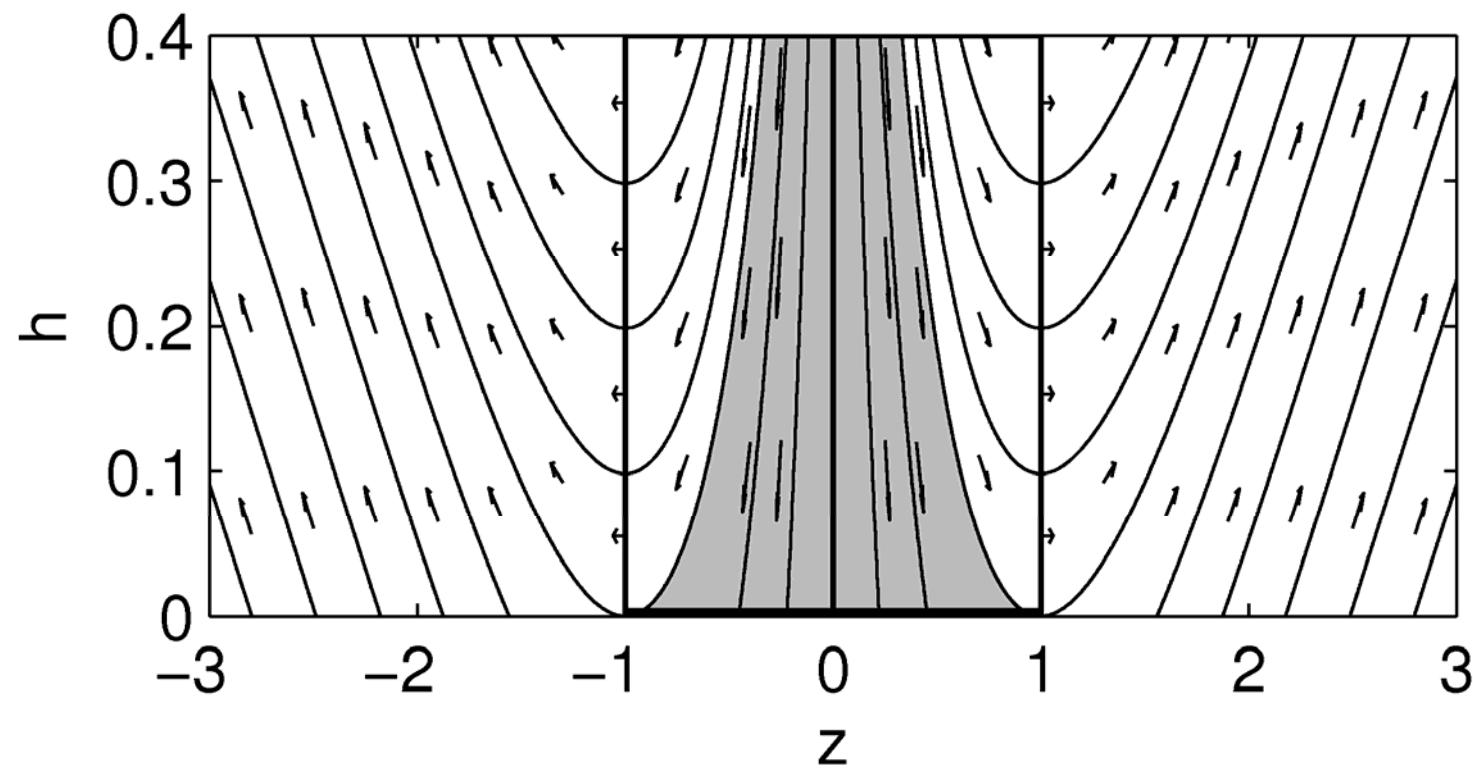


Fig. 2: trajectories