Deep Learning and Physics -- 2019

Deep Random Neural Field

Shun-ichi Amari

RIKEN Center for Brain Science ; Araya

Brief History of AI and NN

First Boom: start

1956 ~ AI neural networks--perceptron

Dartmouth Conf. **Perceptron**

symbol
logic

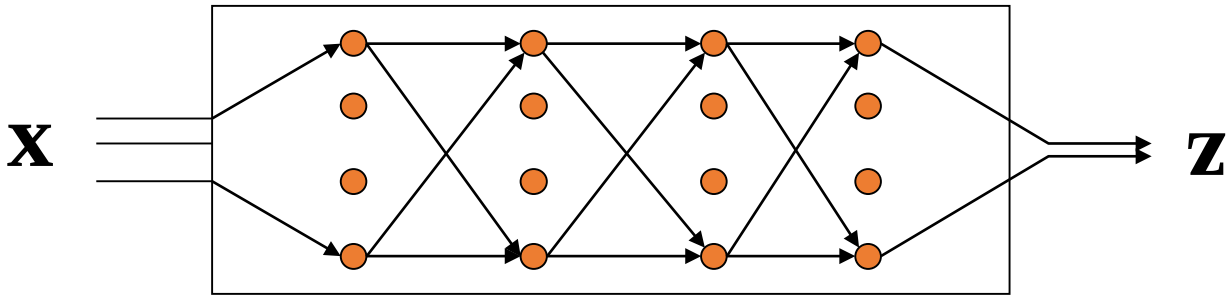
universal computation
learning machine

Dark period (late 1960~1970's)

stochastic gradient descent learning (1967) for MLP

Perceptron

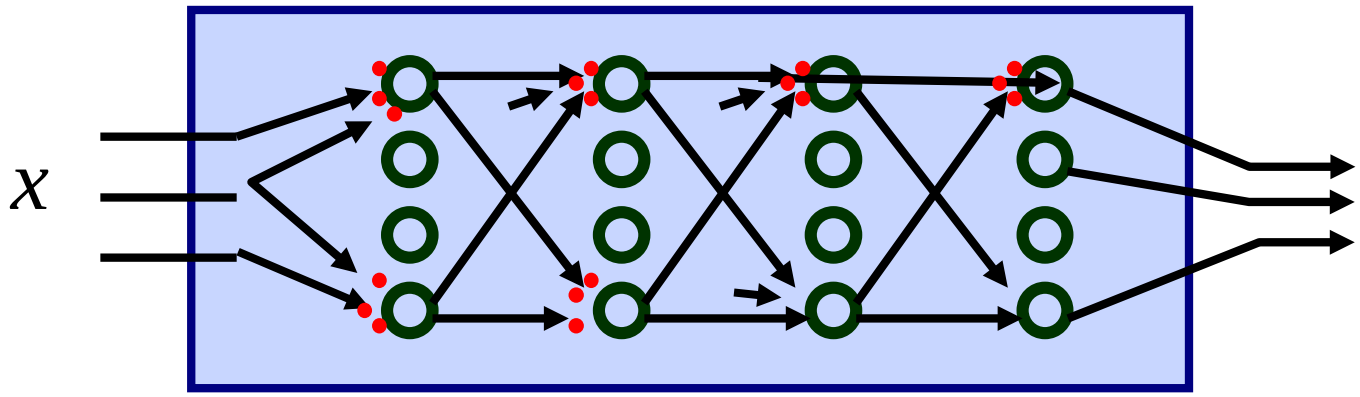
F.Rosenblatt, Principles of Neurodynamics, 1961



McCulloch-Pitts neuron 0,1 binary learning

Multilayer lateral & feedback connection

Deep Neural Networks Rosenblatt: multilayer perceptron



$$z = f(x, W)$$

$$L(\mathbf{x}, W) = \left| y - f(\mathbf{x}, W) \right|^2$$

$$\boldsymbol{w} \to \boldsymbol{w} + \Delta \boldsymbol{w}, \quad \Delta \boldsymbol{w} = -c \frac{\partial L(\mathbf{x}, W)}{\partial W}$$

differentiable : analog neuron

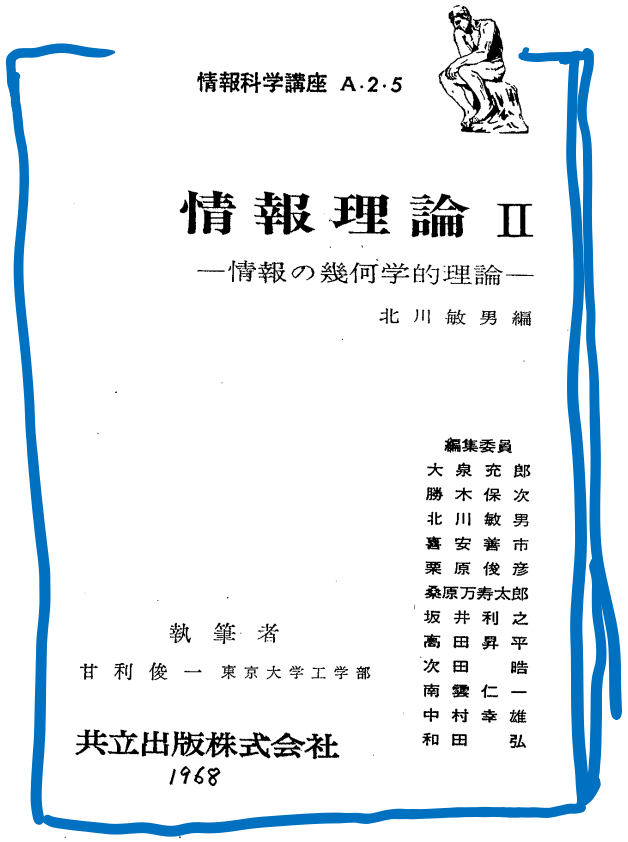
learning of hidden neurons

analog neuron

stochastic gradient learning

Amari, Tsypkin, 1966~67: error back-prop, 1976

First stochastic descent learning of MLP (1967;1968)



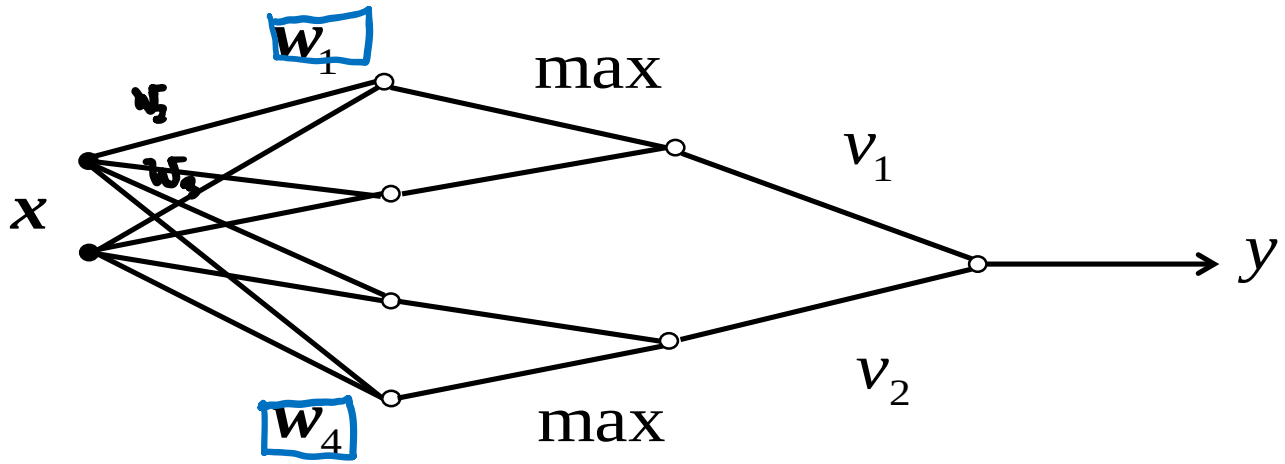
Information Theory II

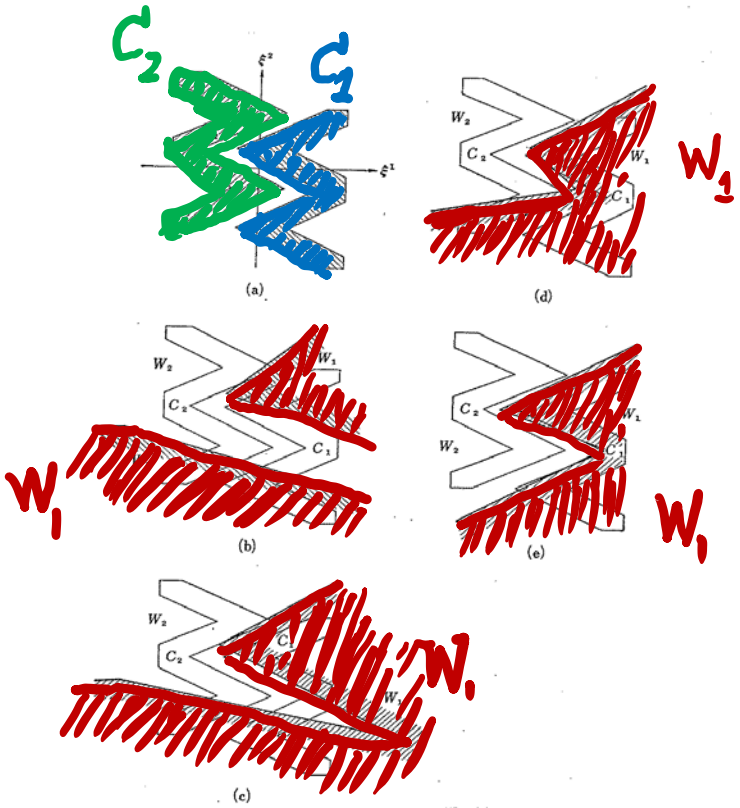
--Geometrical Theory of Information

Shun-ichi Amari
University of Tokyo

Kyoritu Press, Tokyo, 1968

$$f(\boldsymbol{x},\boldsymbol{\theta}) = v_1 \max\{\boldsymbol{w}_1 \cdot \boldsymbol{x}, \boldsymbol{w}_2 \cdot \boldsymbol{x}\} + v_2 \min\{\boldsymbol{w}_3 \cdot \boldsymbol{x}, \boldsymbol{w}_4 \cdot \boldsymbol{x}\}$$




C2
C1
(a)
W2
C2
W1
C1
(d)
W1
W2
C2
C1
(b)
W2
C2
W1
C1
W1
(e)
W2
C2
W1
(c)

Second Boom

1970~ AI

expert system
(MYCIN)

stochastic inference (Bayes)
chess (1997)

1980~ neural networks

MLP (backprop)
associative memory

Third Boom 2010~

Deep learning
Stochastic inference (graphical model; Bayesian; WATSON)

Deep learning
pattern recognition: vision, auditory,
sentence analysis, machine translation
alpha-go

Language processing; sequence and dynamics
(word2vec, deep learning with rec. net)

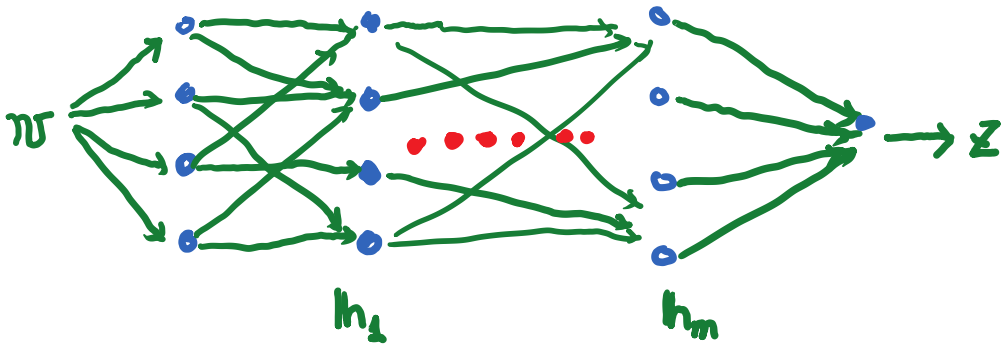
Integration of (symbol, logic) vs (pattern, dynamics)

Deep Learning

Self-Organization + Supervised Learning

RBM: Restricted Boltzmann Machine

Auto-Encoder, Recurrent Net



tricks !!

ideas !

Dropout

Contrastive divergence

Convolution

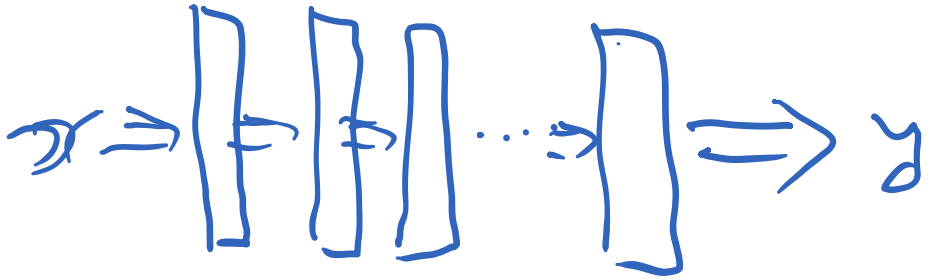
Resnet

ReLU

Adversarial net

Victory of Deep Neural Networks

Hinton 2005, 2006 ~ 2012
many others



visual pattern, auditory pattern

Go-game

sentence analysis, machine translation

adversarial network, pattern generation

Mathematical Neuroscience searches for the principles

mathematical studies using simple idealistic models
(not realistic)

→Computational neuroscience
→AI : technological realization

Mathematical Neuroscience and Brain

Brain has found and implemented the principles

through evolution (random search)
historical restriction
material restriction

Very complex (not smartly designed)

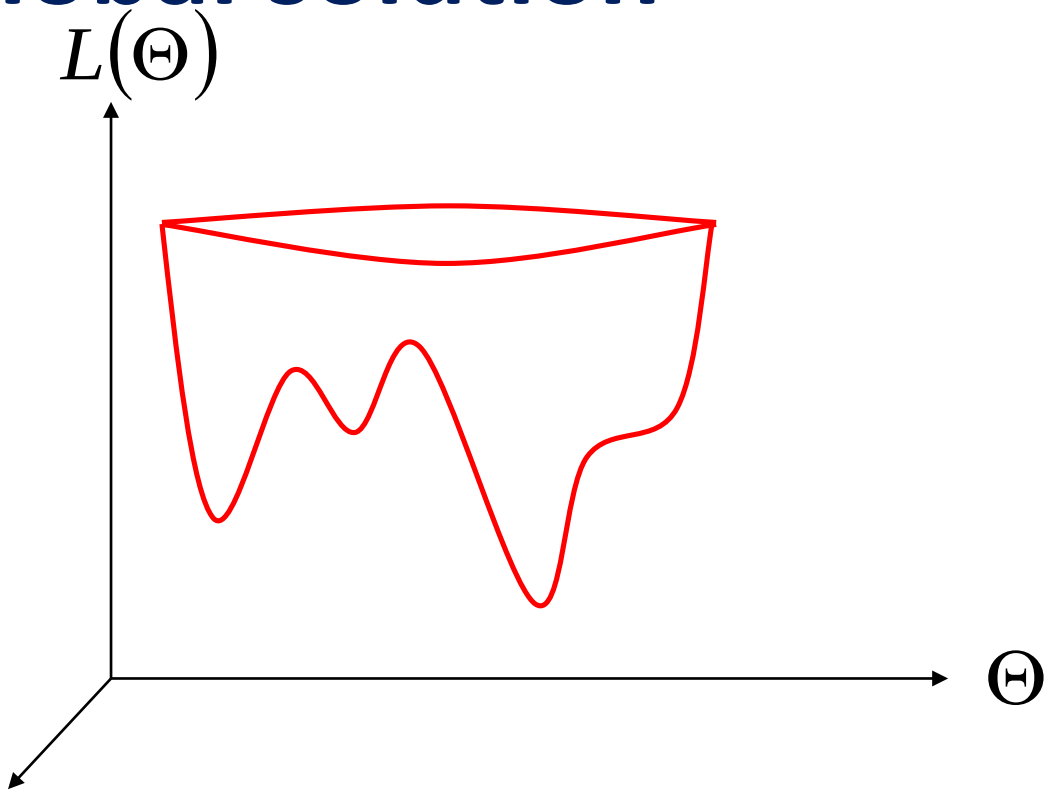
Theoretical Problems on Learning: 1

Local solution and global solution

Simulated annealing

Quantum annealing

:



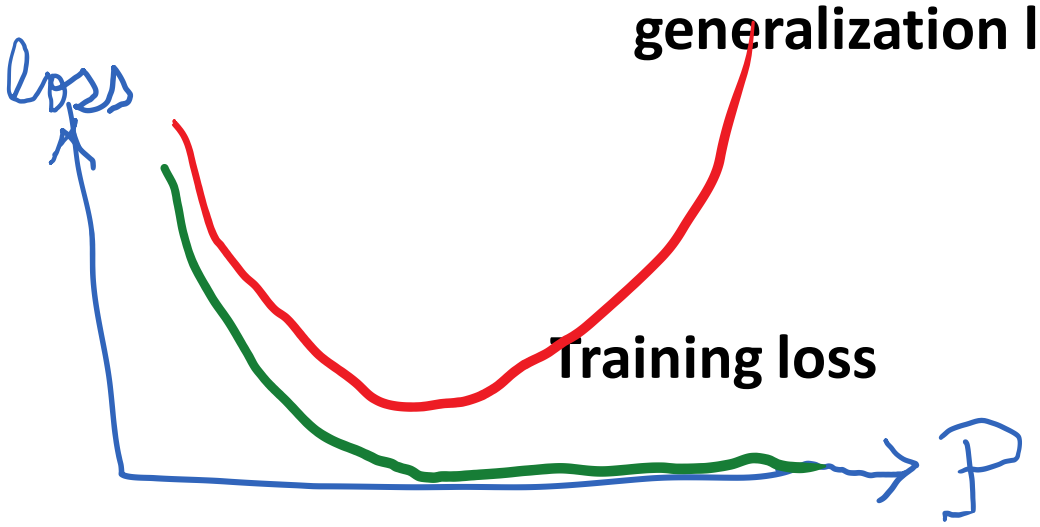
Theoretical Problem of Learning: 2 training loss and generalization loss :overtraining

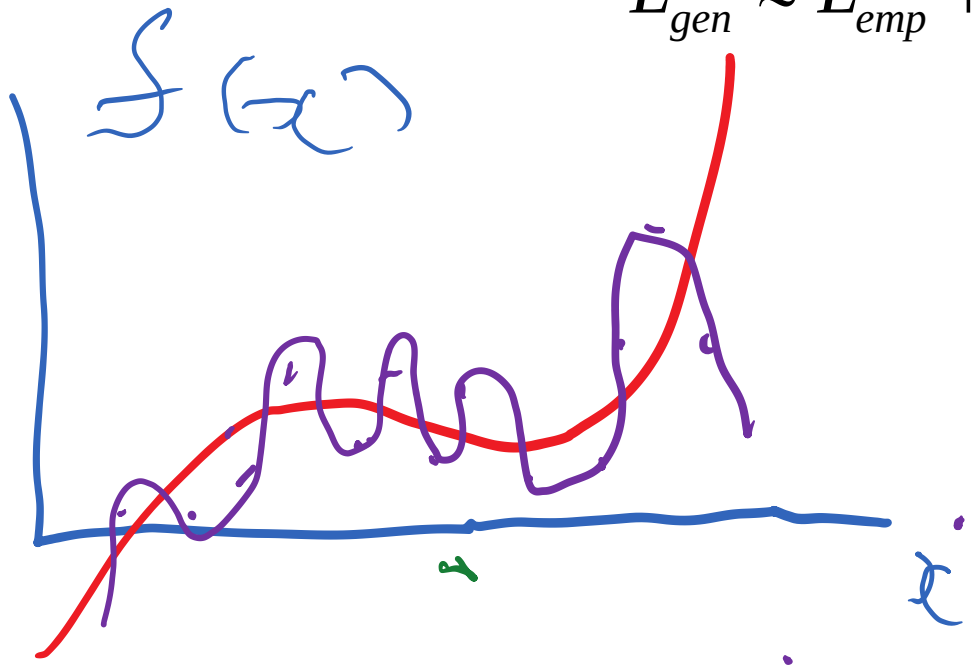
$$y = f(x,\theta) + \varepsilon$$

$$L_{emp} = \frac{1}{N}\sum |\, y_i - f(x_i,\theta)\,|^2$$

$$L_{gen} = E[\,|\, y - f(x,\theta)\,|^2\,]$$

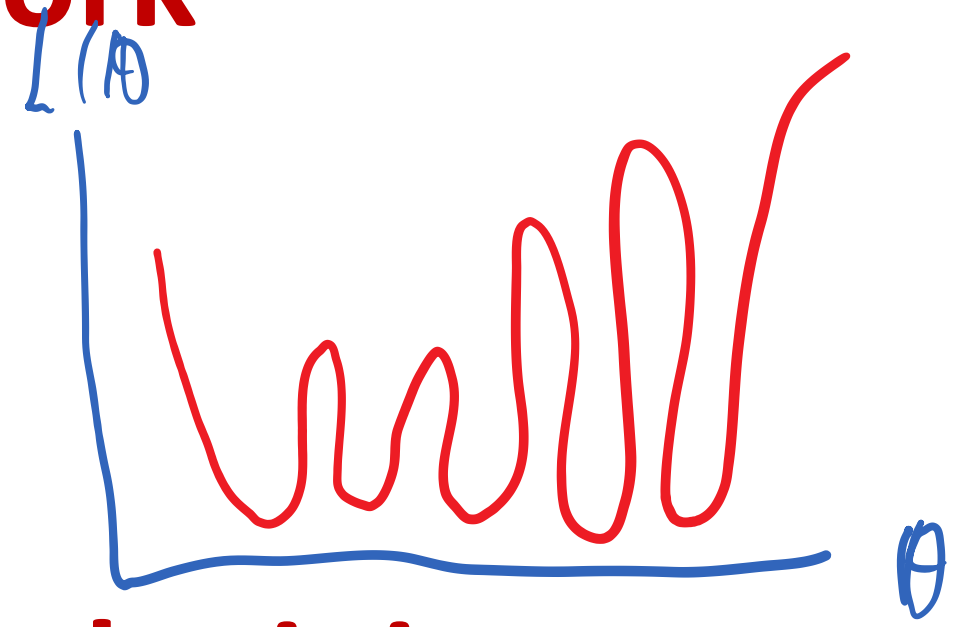
$$L_{gen} \approx L_{emp} + \frac{P}{N}$$





Extremely wdie network

P-> ∞ : P>>N



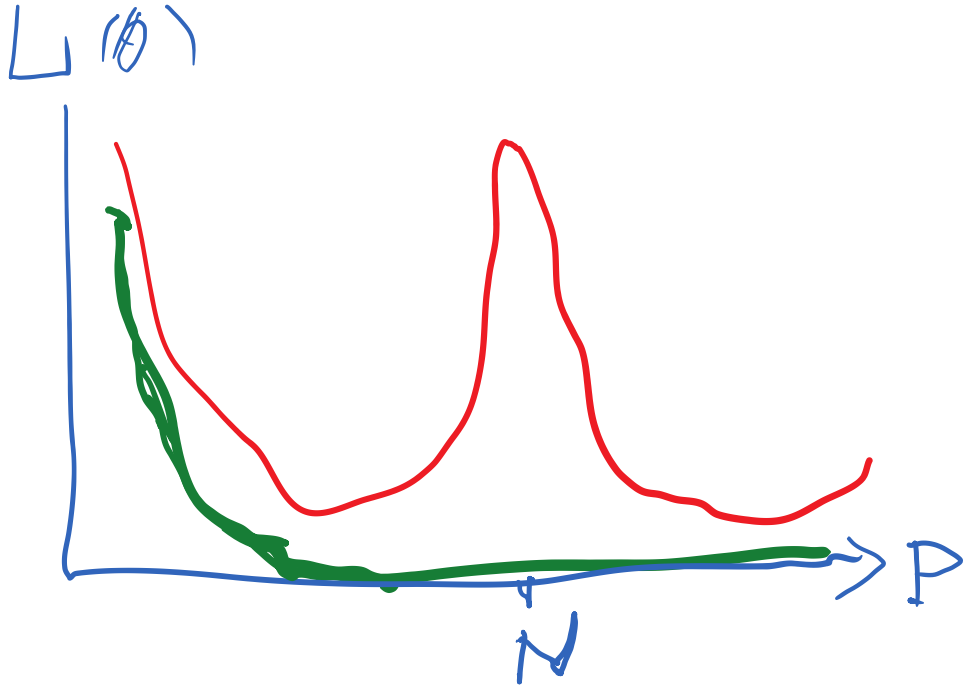
Local minimum =global minimum

Kawaguchi, 2019

Learning curve P>>N

Double descent

Belkin et al. 2019 ; Hastie et al. 2019



Random Neural Network

Random is excellent !!

Random is magic!!

Statistical dynamics

Random code

Random Deep Networks

Poole et. Al., 2016

Schoenholtz et. Al., 2017

~~

Signal propagation

Error back propagation

Jacot et al; Neural tangent kernel

$$y = f(x,\theta); \quad l(x,\theta) = \frac{1}{2}(y - f(x,\theta))^2$$

$$l(x,\theta) = \frac{1}{2}\{y - f(x,\theta)\}^2; \quad e = f(x,\theta) - f(x,\theta^*)$$

ここに数式を入力します。

$$\partial_t \theta = -\eta \partial_\theta l = -\eta \partial_\theta f(x')(f(x,\theta) - f(x,\theta^*))$$

$$\partial_t f(x,\theta) = \partial_\theta f(x) \partial_t \theta = -\eta \partial_\theta f(x) \cdot \partial_\theta f(x') e(x',\theta)$$

K; Gaussian kernel

$$K(x, x'; \theta) = \partial_\theta f(x) \cdot \partial_\theta f(x')$$

$$\partial_t f(x, \theta) = -\eta < K(x, x'; \theta) e(x', \theta) >$$

$K(x, x'; \theta) \approx K_{initial}(x, x')$: Gaussian kernel

$\theta_t \approx \theta_{ini}$

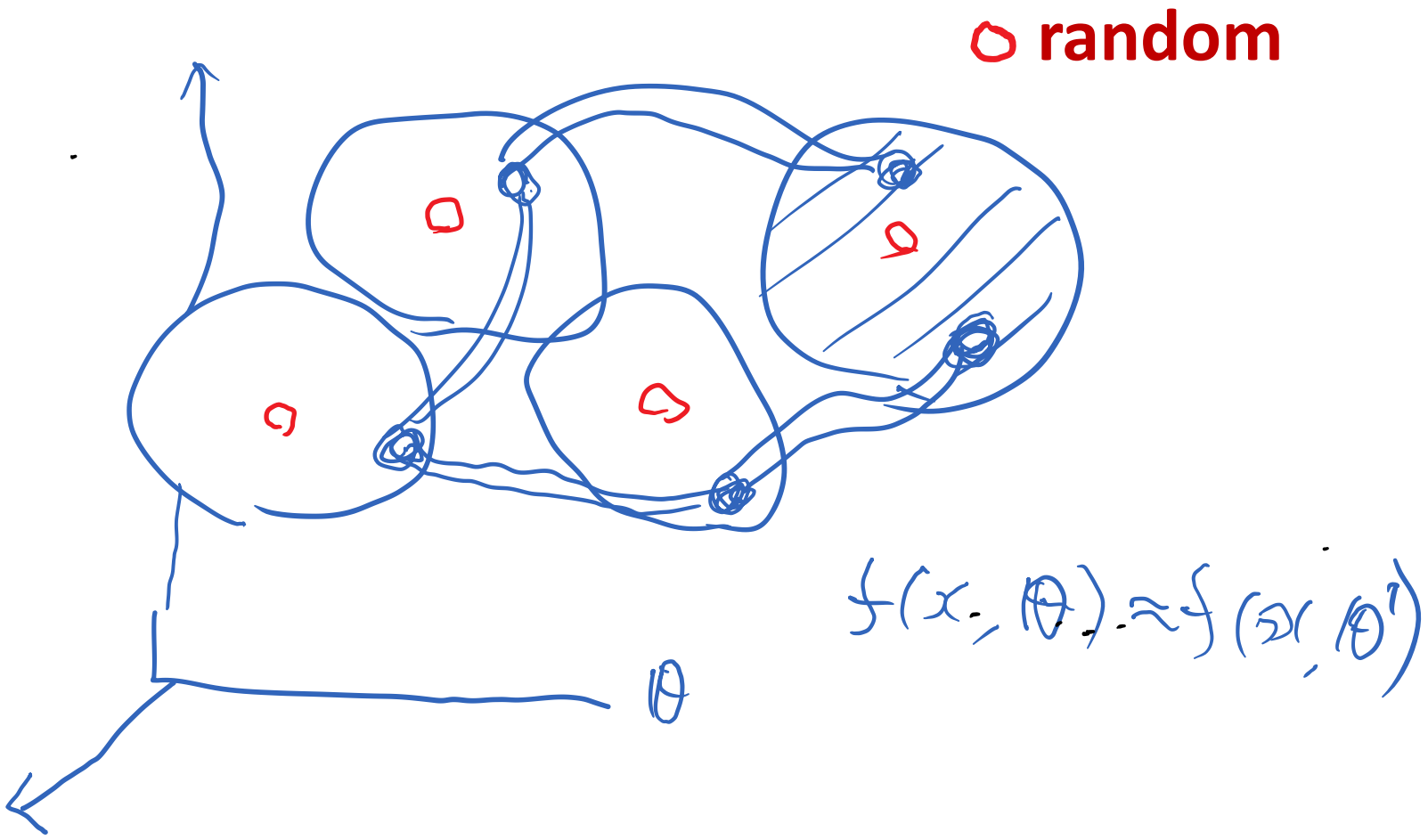
Theorem P>>N

Optimal solution lies near a random network.

Bailey et al
2019

$$w_{ij} = O(\frac{1}{\sqrt{n}})$$

$$\Delta w_{ij} = O(\frac{1}{n})$$



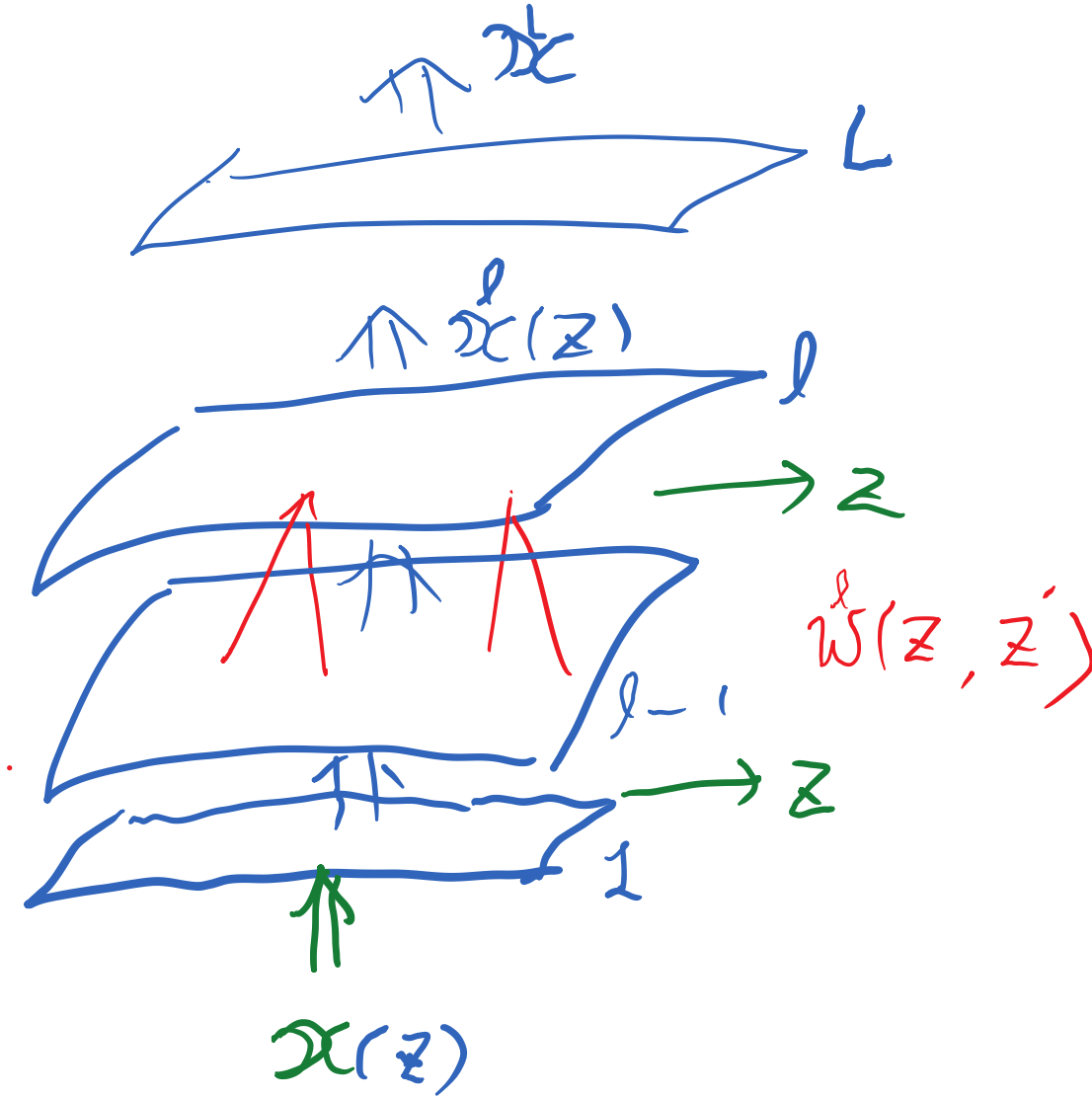
Random Neural Field

$$u(z') = \int w(z, z')\, x^{l-1}(z) dz + b(z')$$

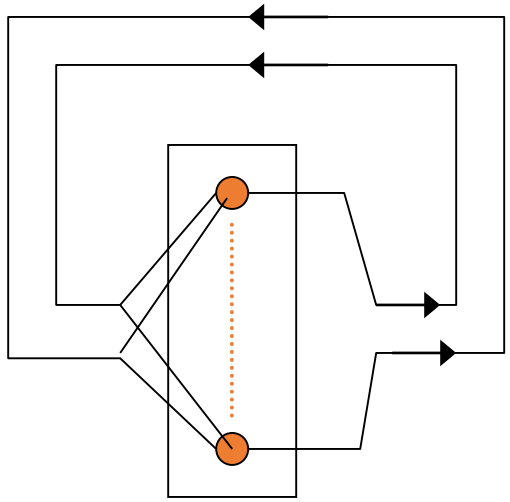
$$x^{l}(z') = \varphi(u(z'))$$

$w(z', z)$: randam

(0 mean Gaussian; correlated)



Statistical Neurodynamics



microdynamics

micro

$$\mathbf{x}(t+1) = T_w \mathbf{x}(t) = \mathrm{sgn}\left(W\mathbf{x}(t)\right)$$

macrodynamics

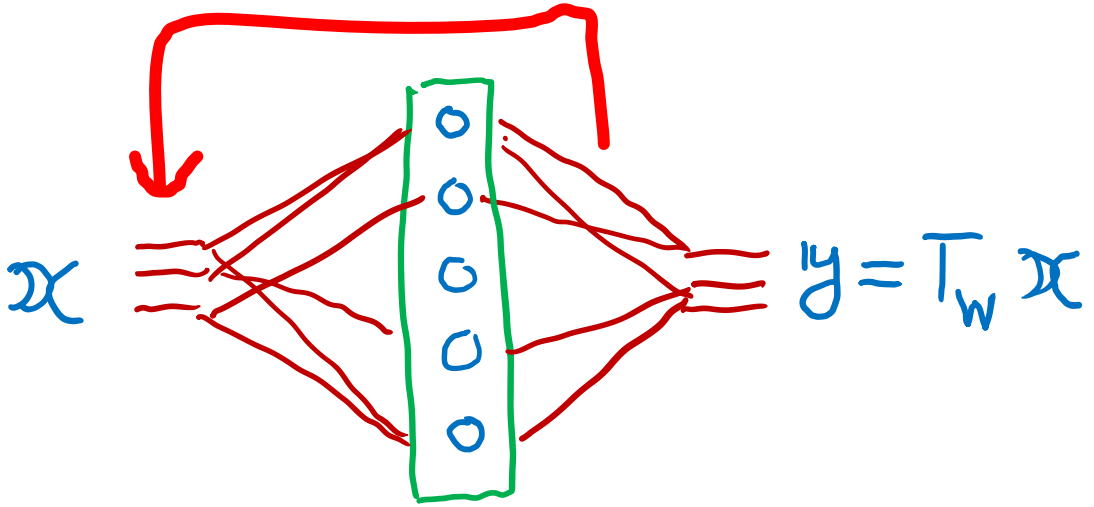
$$X_{t+1} = F(X_t)$$: **macrostate**

$$X_2 = X(\mathbf{x}_2) = X(T_W \mathbf{x}_1)$$

$$X_3 = X(\mathbf{x}_3) = X(T_W T_W \mathbf{x}_1)?$$

Statistical Neurodynamics

Rozonoer (1969)
Amari (1969, 1971; 1973)



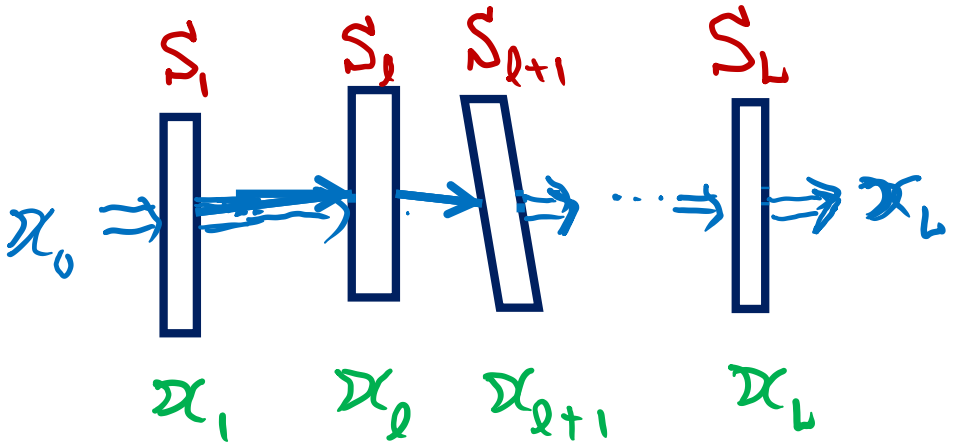
Sompolinski
Amari et al (2013)
Toyoizumi et al (2015)
Poole, …, Ganguli (2016)
Schoenholz et al (2017)
Yang & Schoenholtz (2017),
Karakida, et al (2019)
Jacot et al. (2019) ……

$$w_{ij} \sim N(0,\ 1)$$

Macroscopic behaviors
common to almost all (typical) networks

Random Deep Networks

$$x_i^{l+1} = \varphi\left(\sum_l w_{ij}^l x_j^l + w_{0i}\right)$$



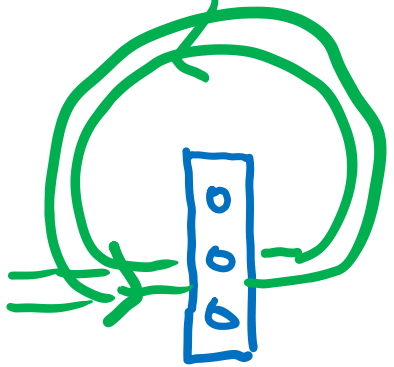
$$A_l = \frac{1}{n_l}\sum_l x_i^2$$

$$A_{l+1} = F(A_l)$$

$$w_{ij} \sim N(0,\ \sigma_l^2 / \sqrt{n})$$

$$w_{0i} \sim N(0,\ \sigma_b^2)$$

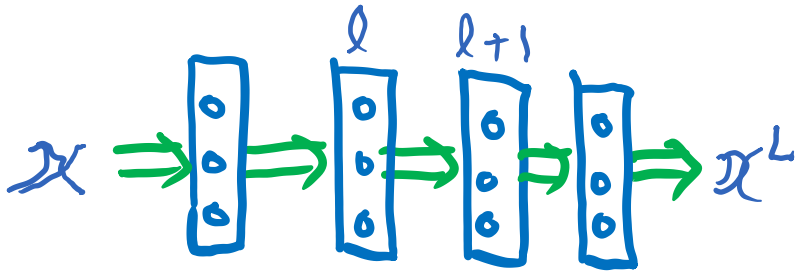
Macroscopic variables



recurrent dynamics

$$\text{activity}: A = \frac{1}{n}\sum x_i^{\,2}$$

$$\text{distance: } D = D[\mathbf{x} : \mathbf{x}']$$



metric, curvature & Fisher information

multilayer dynamics

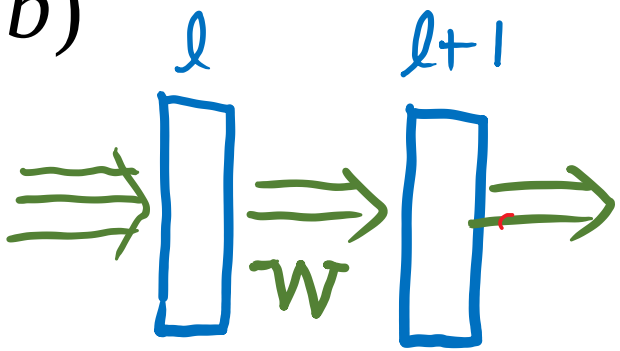
$$A_{l+1} = F(A_l)$$

$$D_{l+1} = K(D_l)$$

Dynamics of Activity: law of large numbers

$$\tilde{x}_i = \varphi(\sum w_{ik} x_k + b_i) = \varphi(u_i) : \tilde{x} = \phi(Wx + b)$$

$$u_i \sim N(0, A)$$



$x \rightarrow \tilde{x} \quad : \quad A \rightarrow \tilde{A}$

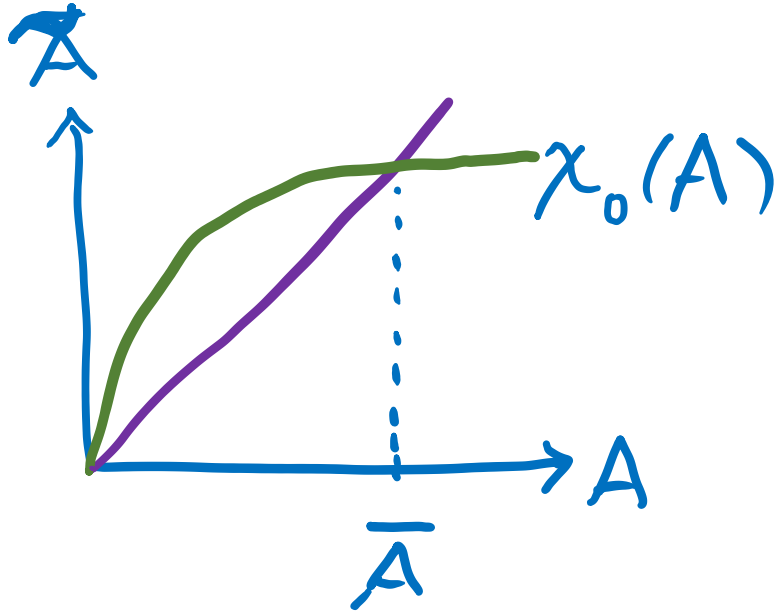
$$\tilde{A} = \frac{1}{n_{l+1}} \sum (\tilde{x}_i)^2 = E[\varphi(u_i)^2] = \chi_0(A)$$

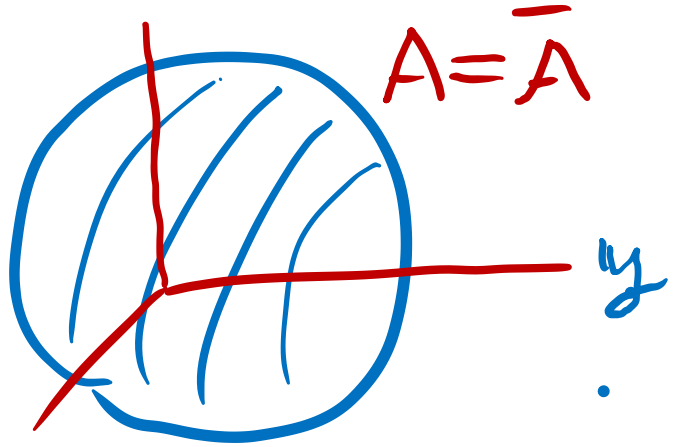
$$\chi_0(A) = \int \varphi^2(\sqrt{A}v) Dv \qquad v \sim N(0,1)$$

$$\chi_0'(0) > 1$$

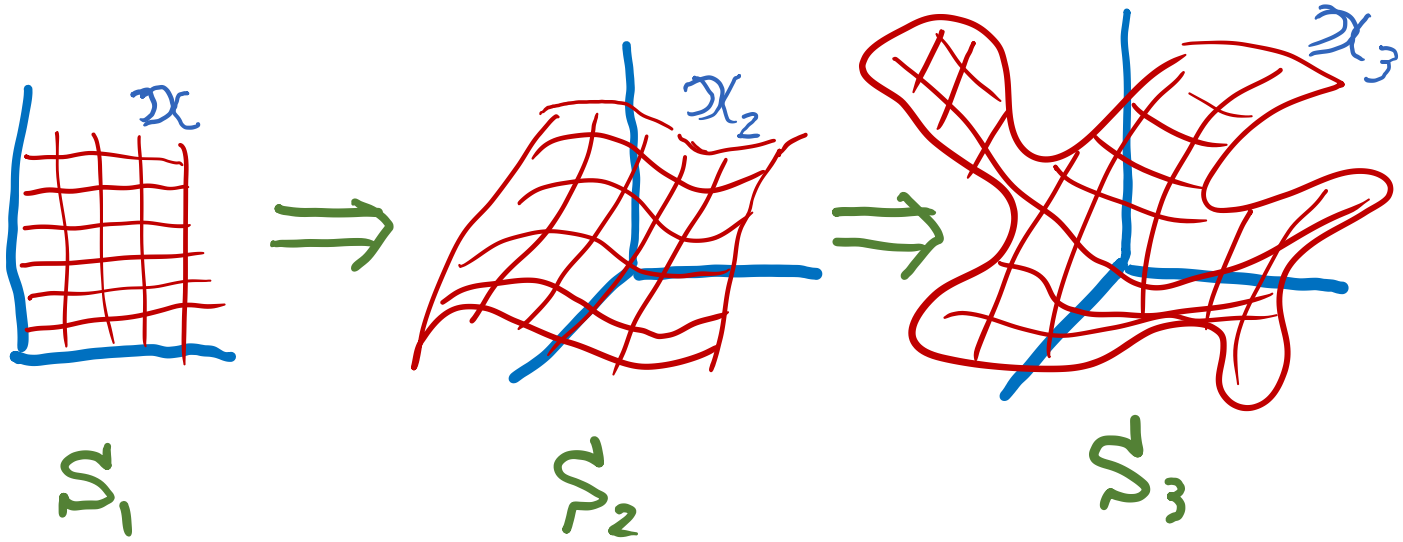
$$\bar{A} = \chi_0(\bar{A})$$

$$\sum x_i^2 \rightarrow \text{converge}$$





Pullback Metric & Curvature



$$\tilde{x} = \phi(Wx)$$

$$ds^2 = \sum^{l} g_{ij}\, dx^i dx^j = \frac{1}{n_l} d\mathbf{x} \cdot d\mathbf{x}$$

Basis vectors

$$\tilde{x} = \phi(Wx)$$

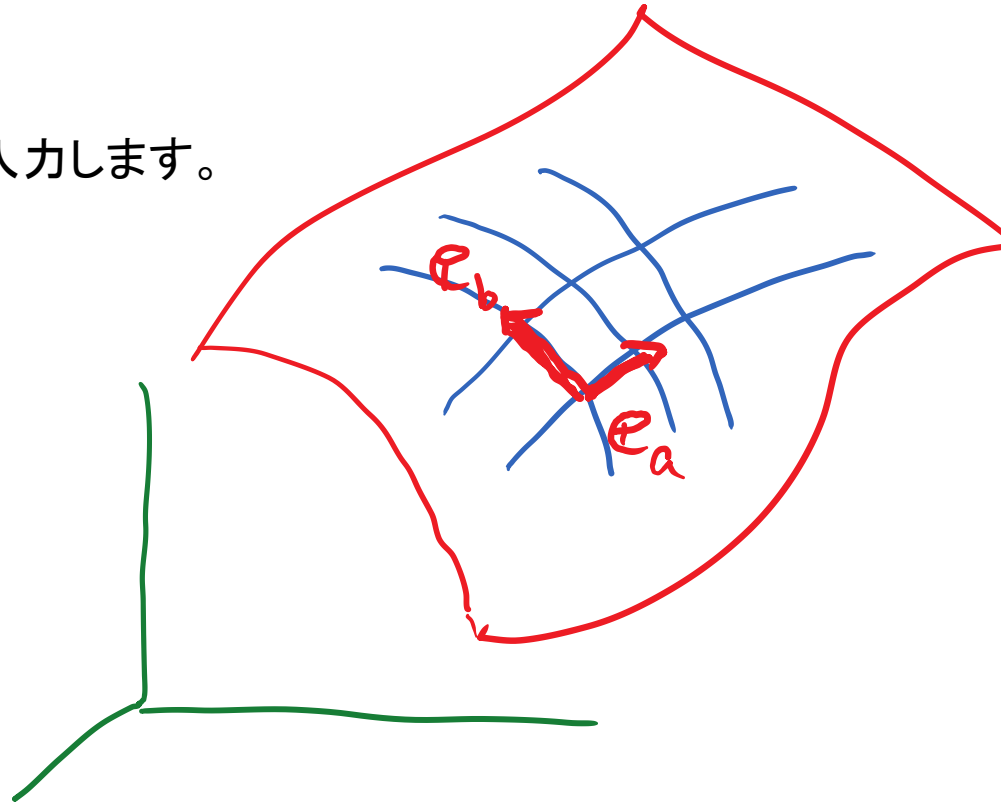
$$d\tilde{x}_{i_l} = \sum \varphi'\left(u_{i_l}\right) W^{i_l}_{i_{l-1}} dx_{i_{l-1}} = \sum B^{i_l}_{i_{l-1}} dx_{i_{l-1}}$$

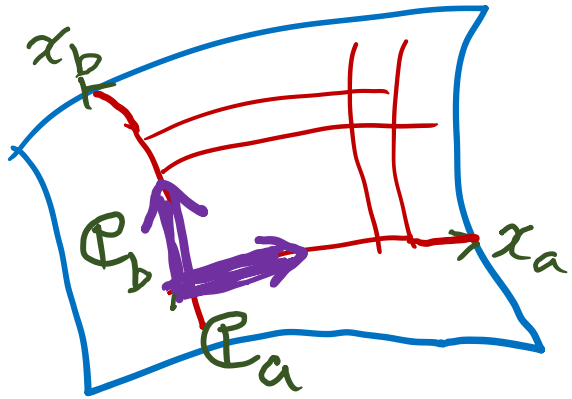
$$d\,\tilde{\boldsymbol{x}}^{l} = B^{l}\, d\,\boldsymbol{x}^{l-1} \qquad \left(= B^{l} \ldots B^{m}\, d\,\boldsymbol{x}^{m-1}\right)$$

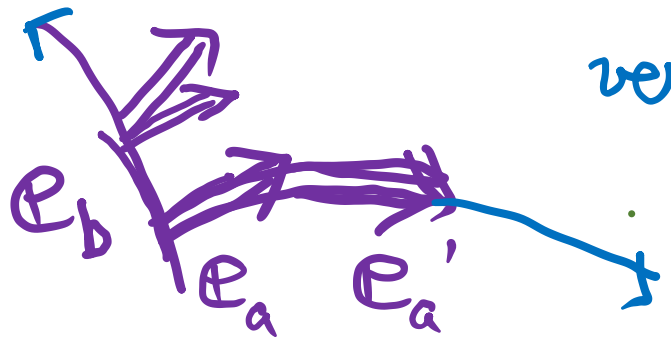
ここに数式を入力します。

$$B^{i_l}_{i_{l-1}} = \varphi'\left(u_{i_l}\right) W^{i_l}_{i_{l-1}} : \text{ Jacobian}$$

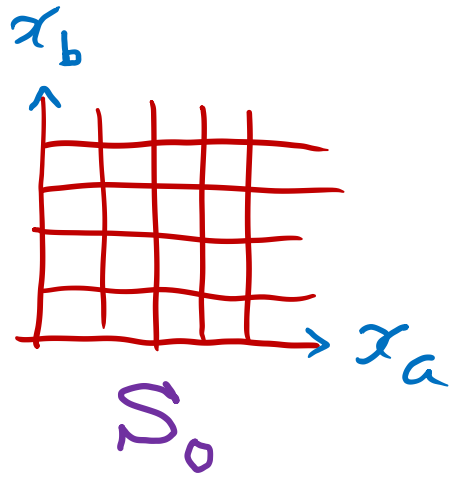
$$\boldsymbol{e}^{l}_{a} = B^{l}\, \boldsymbol{e}^{l-1}_{a} = B^{l} \ldots B^{m}\, \boldsymbol{e}^{m-1}_{a}$$







rotation
vertical



$$g_{ab} = \frac{1}{n_l} \mathbf{e}_a \cdot \mathbf{e}_b$$

Dynamics of Metric

$$d\tilde{x}_a = \sum B_k^a dx_k$$

$$\tilde{\mathbf{e}}_a = B\mathbf{e}_a$$

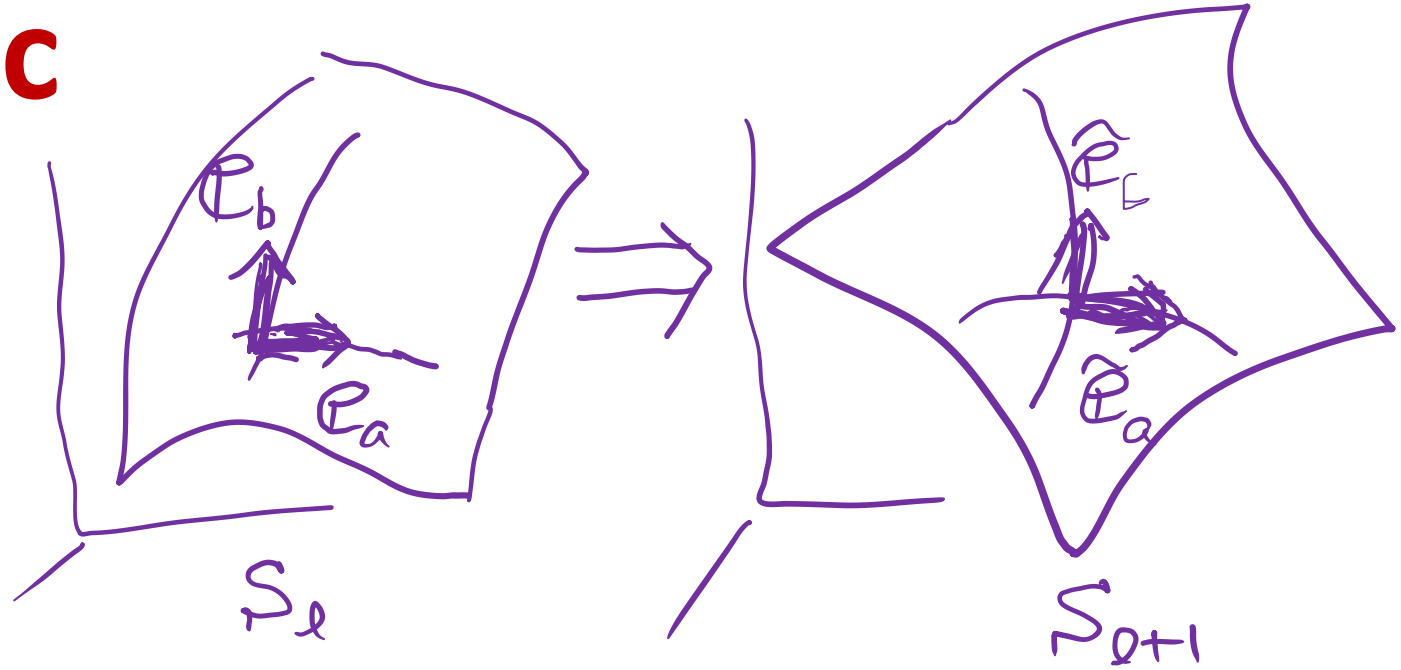
$$B = (B_k^a) = (\varphi'(u_a) w_k^a)$$

$$\tilde{g}_{ab} = \sum B_k^a B_j^b g_{kj}$$

$$\mathrm{E}[\varphi'(u_a))^2 w_k^a w_j^a] = \mathrm{E}[\varphi'(u_a))^2] \mathrm{E}[w_k^a w_j^a]$$

– – mean field approximation

$$\chi_1(A) = \int \varphi'(\sqrt{A}v)^2 Dv$$



$$Dv = \frac{1}{\sqrt{2\pi}} e^{-\frac{v^2}{2}} dv$$

Mectric

Law of large numbers

$$\overset{l}{g}_{ab} = \left\langle \overset{l}{\boldsymbol{e}}_a, \overset{l}{\boldsymbol{e}}_b \right\rangle = BB \; \overset{l-1}{g}_{ab}$$

$$ds^2 = \sum \overset{l}{g}_{ab} \, d \, \overset{l}{x}_a \, d \, \overset{l}{x}_b$$

$$BB = \sum_{i_l} w^{i_l}_{i_{l-1}} \; w^{i_l}_{i'_{l-1}} \varphi'\left(u_{i_l}\right)^2 \approx \sigma_l^2 E\left[\varphi'^2\right] \delta_{i_{l-1} \; i'_{l-1}}$$

$$\chi_1 = \sigma_l^2 E\left[\varphi'\left(u_{i_l}\right)^2\right]$$

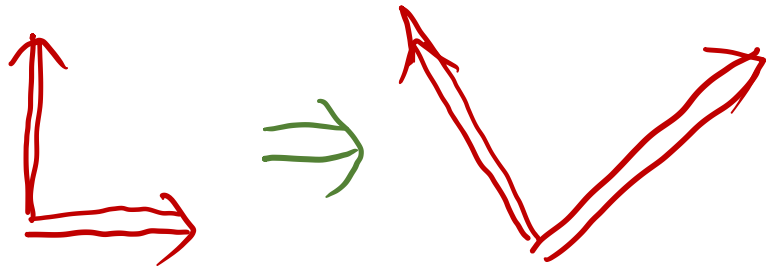
$$\overset{l}{g}_{ij}(x) = (\prod \chi_1(x))g_{ij}(x)$$

conformal geometry

$$\tilde{g}_{ij}(x) = \chi_1(A) g_{ij}(x)$$

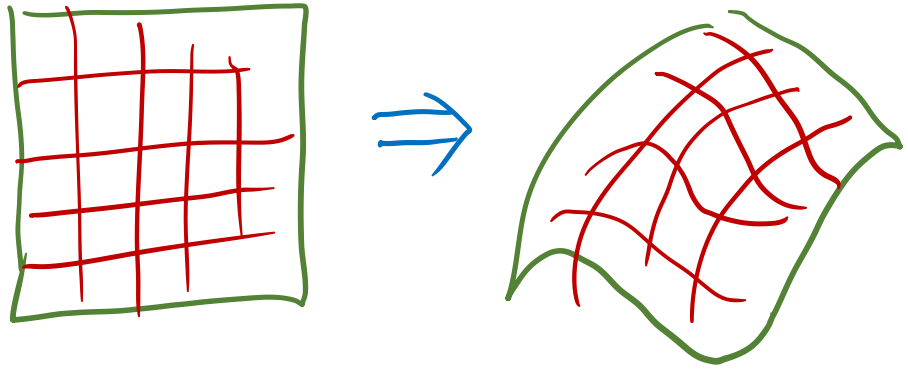
conformal transformation!

rotation, expansion



$$\bar{\chi}_1 = \bar{\chi}_1(\bar{A})\delta_{ij}$$

$$\overset{l}{\Rightarrow} g_{ij} = \bar{\chi}_1^l \delta_{ij}$$



Domino Theorem

$$\frac{\partial \overset{l}{\boldsymbol{x}}}{\partial \overset{m}{\boldsymbol{x}}} = B\frac{\partial \overset{l-1}{\boldsymbol{x}}}{\partial \overset{m}{\boldsymbol{x}}} = BB\frac{\partial \overset{l-2}{\boldsymbol{x}}}{\partial \overset{m}{\boldsymbol{x}}} = BBB\cdots$$

$$\frac{\partial \overset{l}{\boldsymbol{x}}}{\partial \overset{m}{W}} = B\frac{\partial \overset{l-1}{\boldsymbol{x}}}{\partial \overset{m}{W}} = BB\frac{\partial \overset{l-2}{\boldsymbol{x}}}{\partial \overset{m}{W}} = BBB\cdots$$

$$\Sigma\delta_{i_L i'_L} B^{i_L}_{i_{L-1}} B^{i'_L}_{i'_{L-1}} = \chi_1 \delta_{i_{L-1} i'_{L-1}}$$

$$\Sigma\delta_{i_L i'_L} B^{i_L}_{i_{L-1}} B^{i'_L}_{i'_{L-1}} BBBB = \chi_1\chi_1\chi_1\chi_1 \delta_{i_{L-1} i'_{L-1}}$$

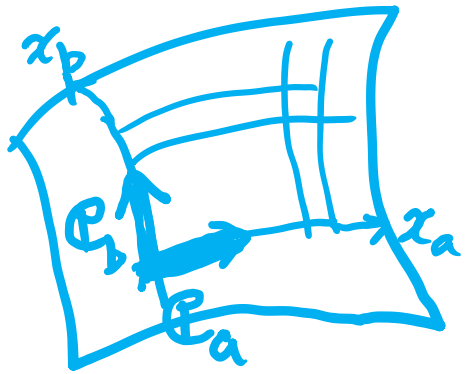
Dynamics of Curvature

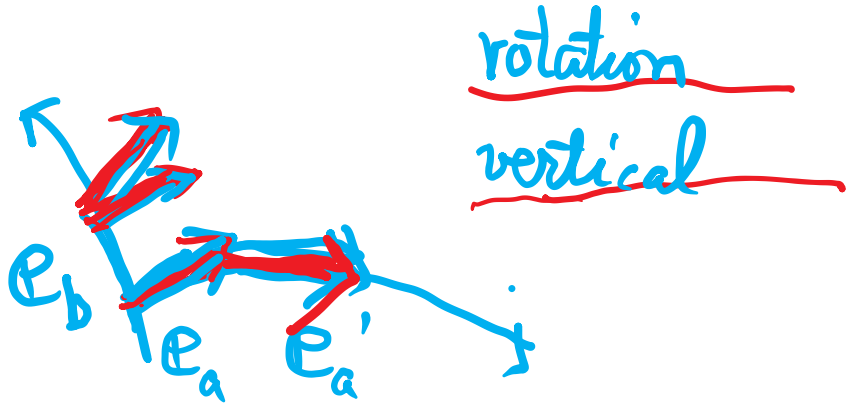
$$\tilde{H}^i_{ab} = \nabla_a \tilde{\mathbf{e}}_b = \partial_a \partial_b \tilde{x}^i$$

$$= \varphi''(u_i)(\mathbf{w}\cdot\mathbf{e}_a)(\mathbf{w}\cdot\mathbf{e}_b) + \varphi'(\mathbf{w}\cdot\partial_a \mathbf{e}_b)$$

$$\tilde{\mathbf{H}}_{ab} = \mathbf{H}^{\perp}_{ab} + \mathbf{H}^{\parallel}_{ab}$$

$$\tilde{H}^2_{ab} = |\tilde{\mathbf{H}}_{ab}|^2$$





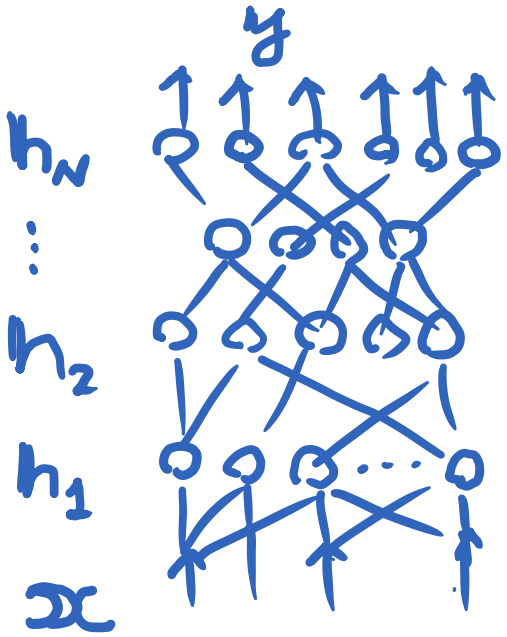
$$\chi_2(A) = \int \varphi''(\sqrt{A}v)^2 Dv$$

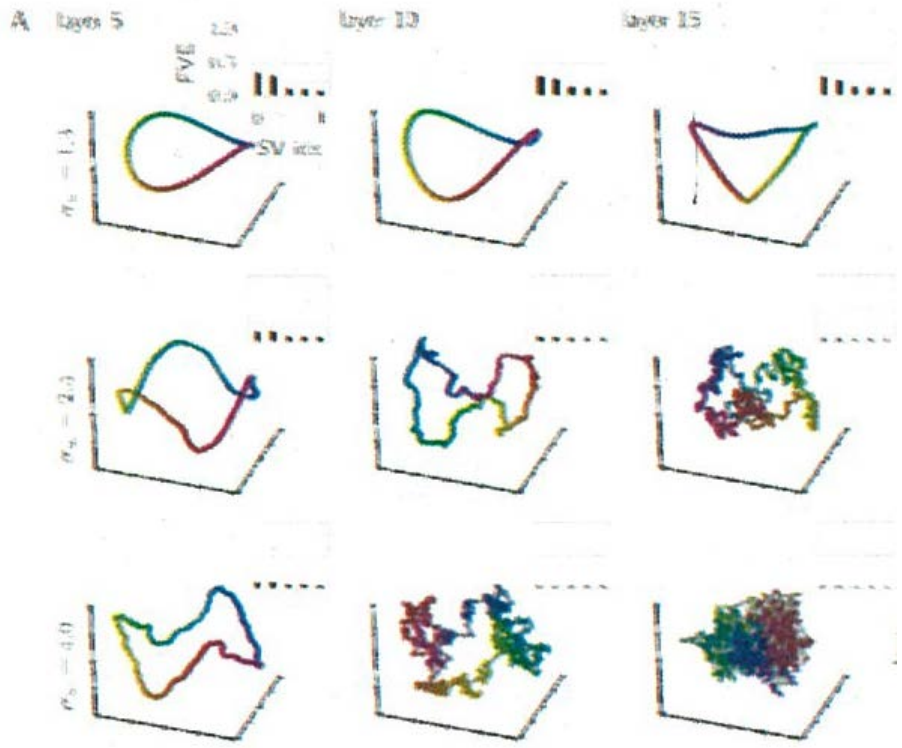
$$H_{ab}^{l+1\,2} = \frac{1}{n\chi_1^2}\chi_2(A^l)(2\delta_{ab}+1) + \chi_1(A^l)H_{ab}^{l\,2}$$

$$\chi_1 > 1$$

exponential expansion! creation is small!

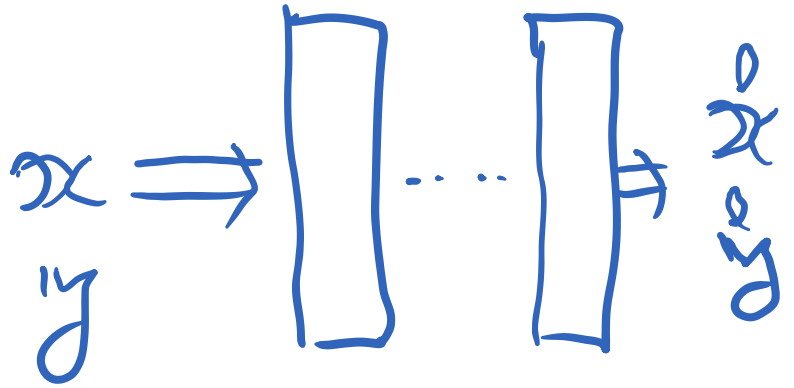
Poole et al (2016)
Deep neural networks





Distance

$$D[x, y] = \frac{1}{n}\sum |x_i - y_i|^2$$

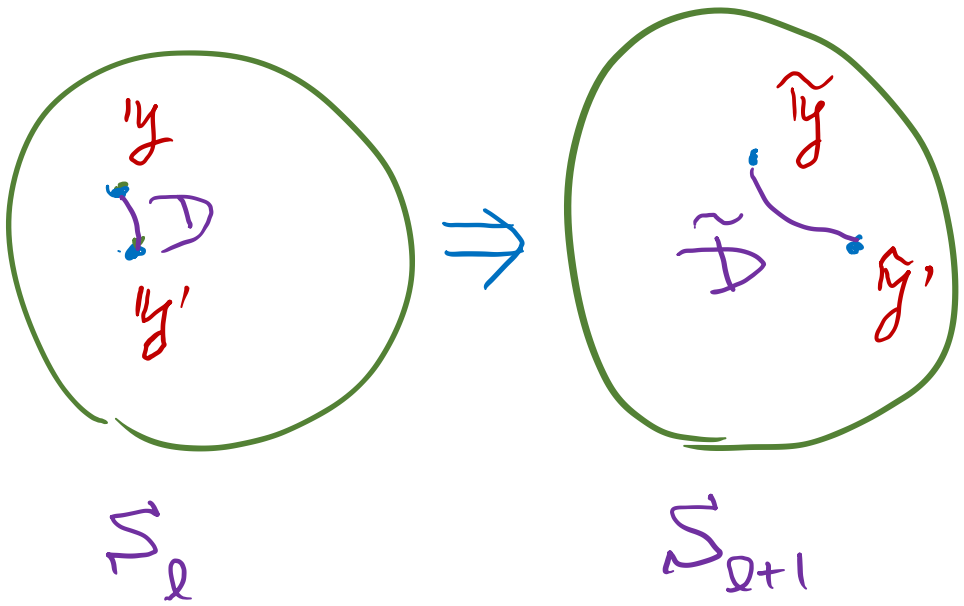


Dynamics of Distance (Amari, 1974)

$$D(x,x') = \frac{1}{n}\sum (x_i - x_i')^2$$

$$C(x,x') = \frac{1}{n} x \cdot x' = \sum x_i x_i'$$

$$D = A + A' - 2C$$



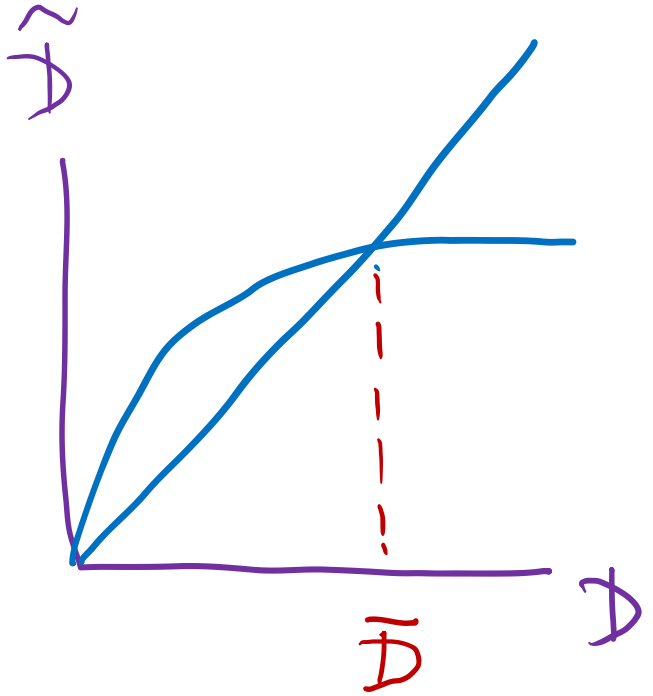
$$u_i = \sum w_k^i y_k \qquad \sim N(0, V)$$

$$u'_i = \sum w_k^i y'_k \qquad V = \begin{pmatrix} A & C \\ C & A' \end{pmatrix}$$

$$\tilde{C} = \mathrm{E}[\varphi(\sqrt{A-C}\,\varepsilon + \sqrt{C}\nu)\varphi(\sqrt{A'-C}\,\varepsilon + \sqrt{C}\nu)]$$

$$D_{l+1} = K(D_l)$$

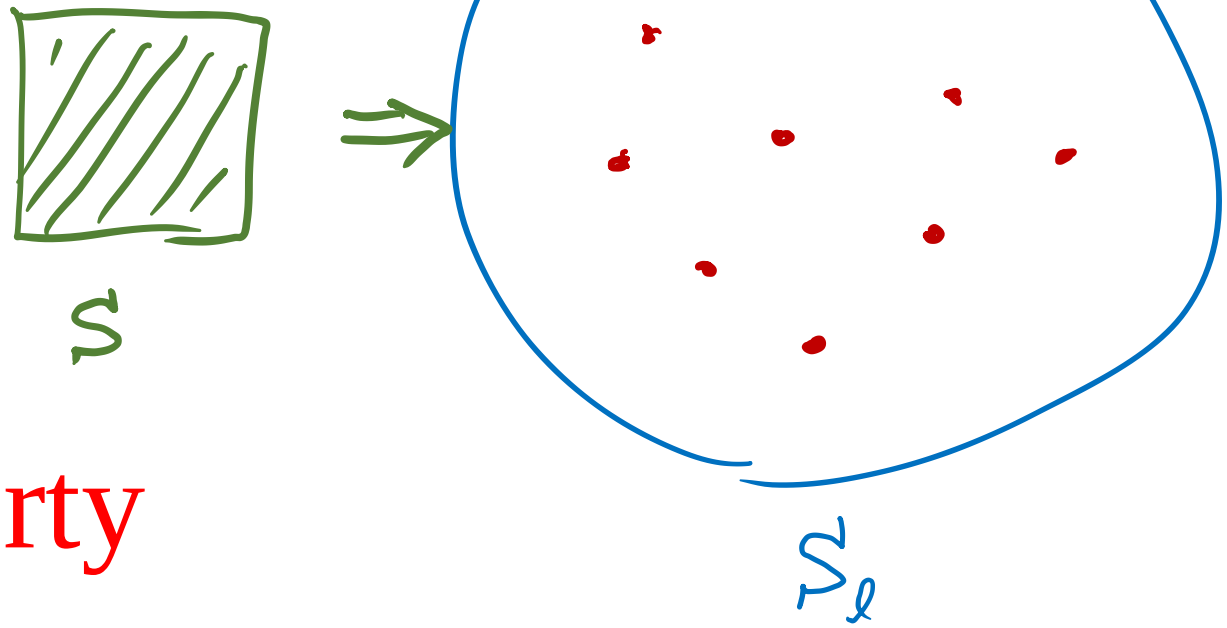
$$\frac{d\tilde{D}}{dD} = \chi_1 > 1$$



Problem!

$$D(\mathbf{x}_l, \mathbf{x}'_l) \to \bar{D}: \quad l \to \infty$$

$$\bar{D} = K(\bar{D})$$



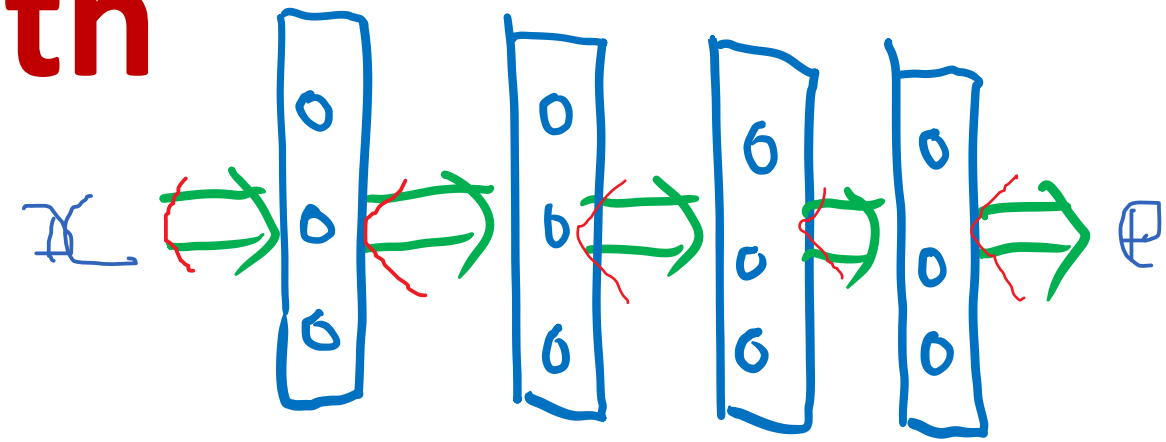
equi-distance property

dynamics of distance

$$\lim_{\substack{n\to\infty \\ L\to\infty}} \overset{L}{D}(\overset{L}{x}, y)$$

$$\lim_{n\to\infty}\lim_{L\to\infty} \overset{L}{D}(\overset{L}{x}, y) \neq \lim_{L\to\infty}\lim_{n\to\infty} \overset{L}{D}(\overset{L}{x}, y)$$

Feedback Path



Error backprop

Fisher Information

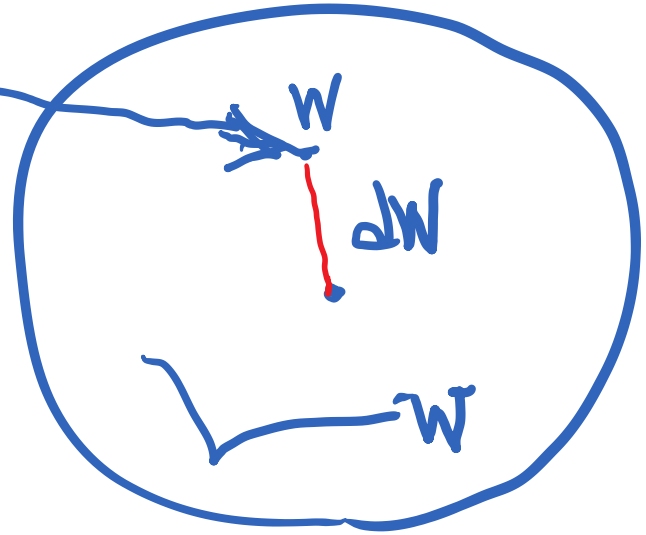
Stochastic model : parameter space
manifold of probability distributions

$$y = \varphi(Wx\varphi(Wx.....)..) + \varepsilon; \qquad \varepsilon \sim N(0,1)$$

$$p(y, x : W) = c\exp\{-\frac{1}{2}(y - \varphi(x;W))^2\}q(x)$$

$$G = E_x[\nabla_W \log p(y, x : W) \nabla_W \log p(y, x : W)]$$

$$ds^2 = dWGdW$$

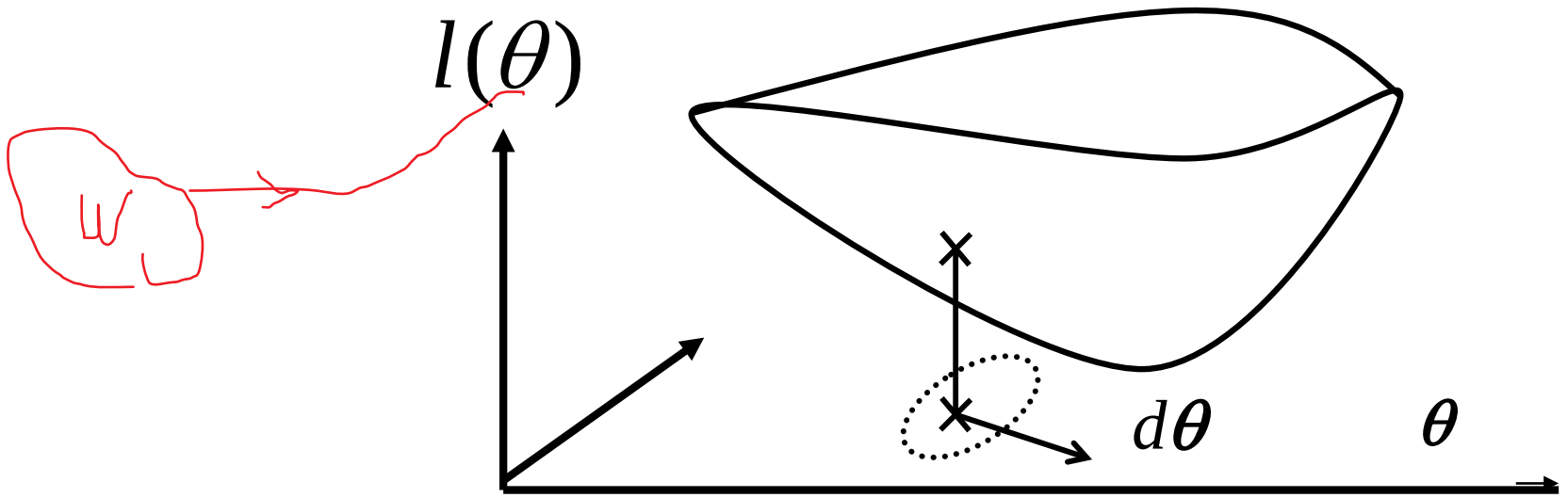


Riemannian

Fisher information

error

Learning: stochastic gradient descent
Steepest Direction---Natural Gradient



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

$$\nabla l = \left(\frac{\partial l}{\partial \theta_1}, \cdots, \frac{\partial l}{\partial \theta_n} \right)$$

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

Natural Gradient

$$\max \qquad dl = l(\boldsymbol{\theta} + d\boldsymbol{\theta}) - l(\boldsymbol{\theta})$$

$$\left|d\boldsymbol{\theta}\right|^2 = \varepsilon \qquad \mathrm{KL}[\mathrm{p}(\mathrm{x},\theta):\mathrm{p}(\mathrm{x},\theta + d\theta)] = \varepsilon$$

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

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

Information Geometry of MLP

Natural Gradient Learning :

S. Amari ; H.Y. Park

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

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

Fisher Information

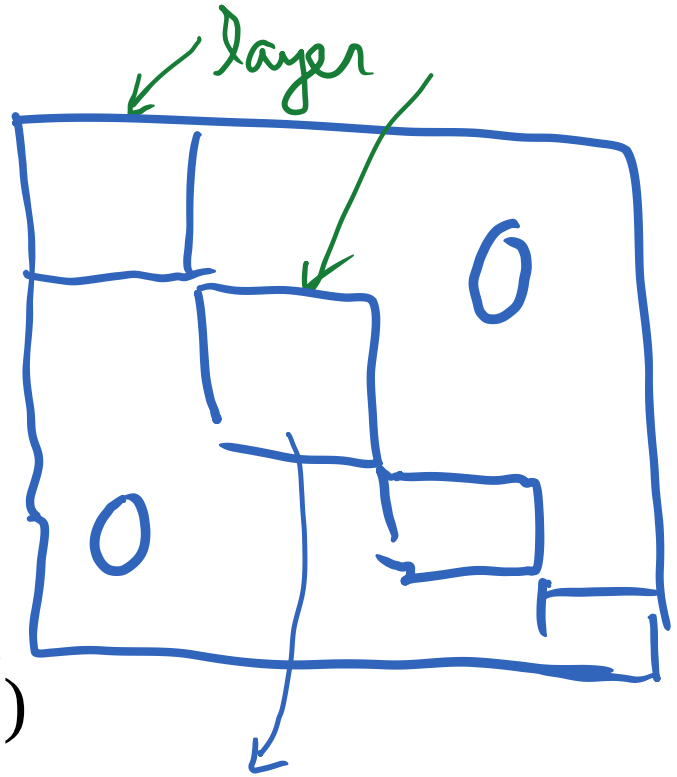
$$G = E_x\left[\frac{\partial\varphi}{\partial W_m}\frac{\partial\varphi}{\partial W_l}\right]$$

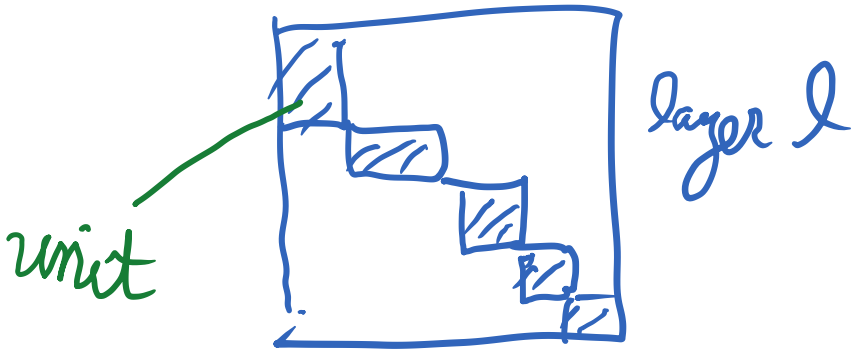
$$\frac{\partial\varphi^l}{\partial W_m} = \varphi' W\frac{\partial\varphi^{l-1}}{\partial W_m} = B\frac{\partial\varphi^{l-1}}{\partial W_m} = BB\ldots B\frac{\partial\varphi^{m+1}}{\partial W_m}$$

$$G(W_l, W_m) = \prod \chi_1 E_x\left[\varphi'\left(\overset{l}{\boldsymbol{w}}_i\right)^2 \overset{l-1}{\boldsymbol{x}}\ \overset{l-1}{\boldsymbol{x}}\right] + O_p(1/\sqrt{n})$$

$$G(W_l, W_m) = 0 \sim O_p(1/\sqrt{n}), \qquad l \neq m$$

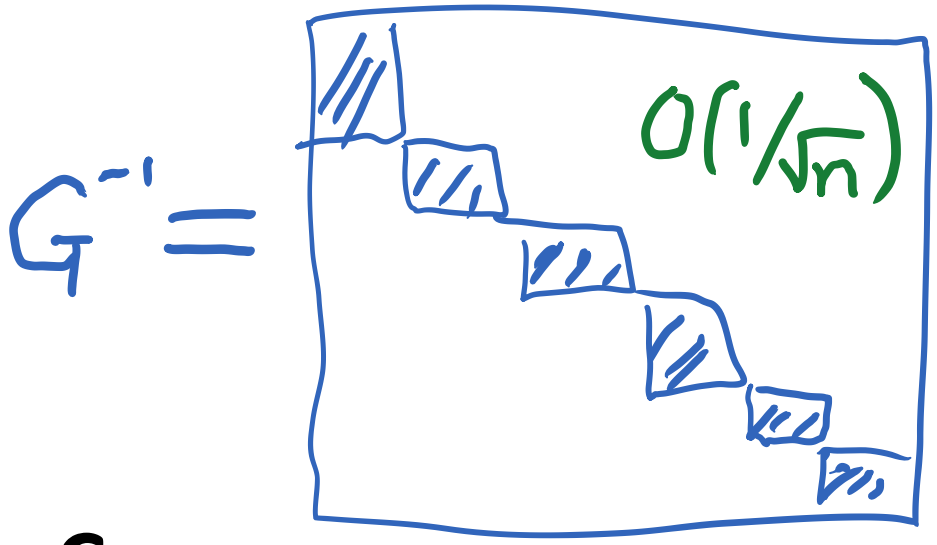
$$G\left(\overset{l}{\boldsymbol{w}}_i, \overset{l}{\boldsymbol{w}}_j\right) = 0 \sim O_p(1/\sqrt{n}), \quad i \neq j$$



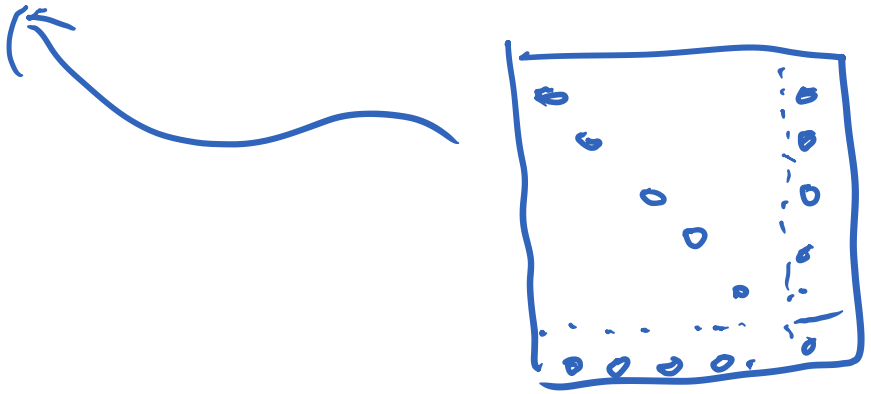


Unitwise natural gradient

$$\Delta W = -\eta G^{-1} \nabla_W l$$



Y. Ollivier; Marceau-Caron



Goodnews and bad news

G* : unitwise-diagonal matrix

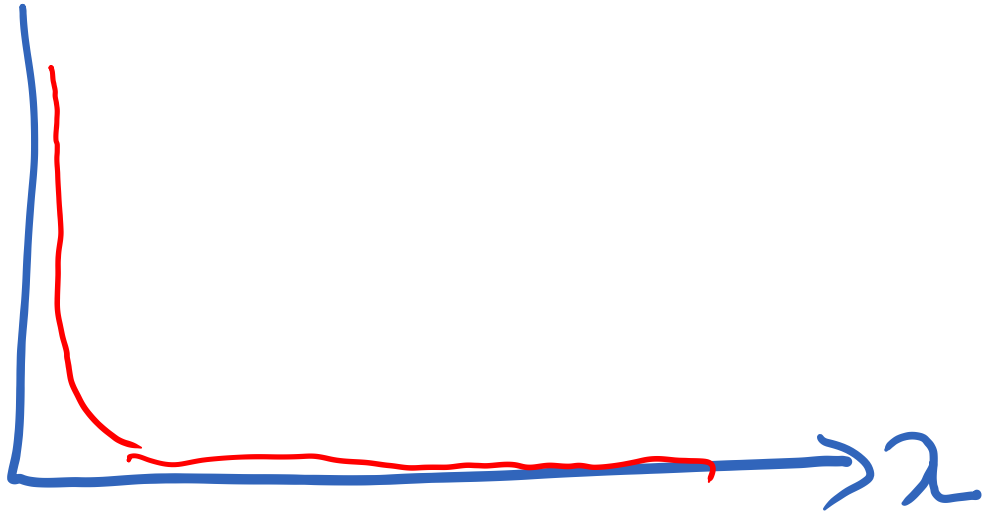
$$G \to G^* : \; n \to \infty$$

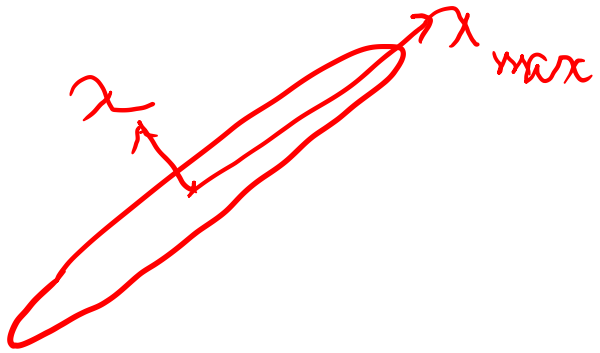
$$G^{-1} \not\to G^{*-1} : \; n \to \infty$$

Karakida theory

eigenvalues of G

$$\sum \lambda_i = \frac{1}{n}, \qquad \sum \lambda_i^2 = O(1)$$





distorted Riemannian metric

G^2

References:

Poole, …, Ganguli (2016)
Schoenholz et al (2017)
Yang & Schoenholtz (2017), ……

S. Amari, R. Karakida & M. Oizumi, Statistical neurodynamics of deep networks: Geometry of Signal Spaces. arXiv:1808.07169v1, 2018.

S. Amari, R. Karakida & M. Oizumi, Fisher information and natural gradient learning of random deep networks. arXiv:1808.07172v1, 2018 (AISTATS-19).

R. Karakida, S. Akaho & S. Amari, Universal statistics of Fisher information in deep neural networks: Mean field approach. arXiv: 1806.01316, 2018 (AISTATS-19).