IEICE ICT Pioneers Webinarシリーズ

数理工学から見たICT

—— 情報幾何学と人工知能（深層学習）
歴史的発展、現状、将来への希望

甘利俊一

理化学研究所　　栄誉研究員；東京大学名誉教授

数理工学への道：幸運な我が人生

数理工学への進学（1956年）
九州大学（1963年）

電子情報通信学会

数理脳科学
情報幾何、確率推論

東大から理研へ

趣味としての研究と囲碁

数理工学への道

数理工学とは方法論である

回路のトポロジー
連続体力学と非リーマン幾何
情報と幾何

学習とパターン認識
神経回路網モデルーパーセプトロン、
数理脳科学

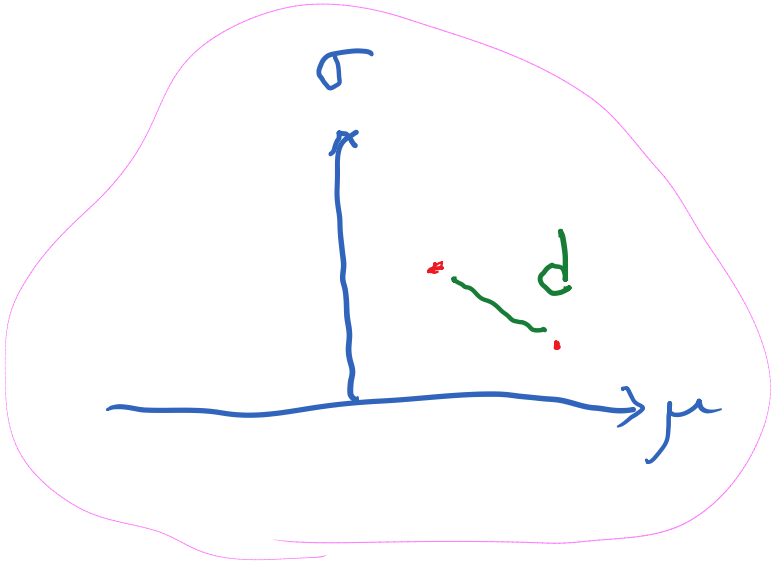
情報幾何
ICA（独立成分解析）
Wasserstein 距離
ランダム神経場

情報幾何の事始め　情報と幾何

確率分布族の作る空間
　ガウス分布

$$p(x;\mu,\sigma)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

負の定曲率空間
（非ユークリッド幾何、ポアンカレの半平面）



学位論文：通信の幾何学理論一本会論文賞

Rao, 1945；Hotelling, 1929

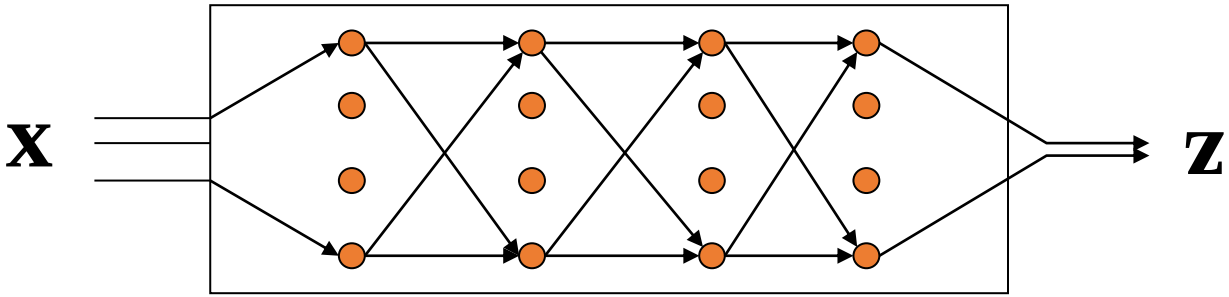
機械学習事始め

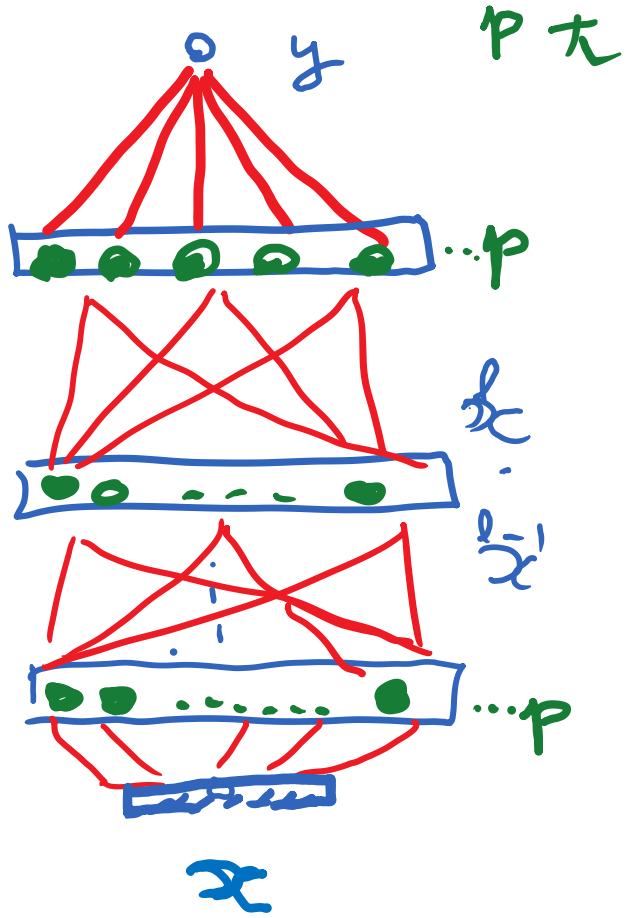
人工知能
記号と論理――学習神経回路網

パーセプトロン：　Rosenblatt
数学者と研究会

Perceptron

F.Rosenblatt, Principles of Neurodynamics, 1961



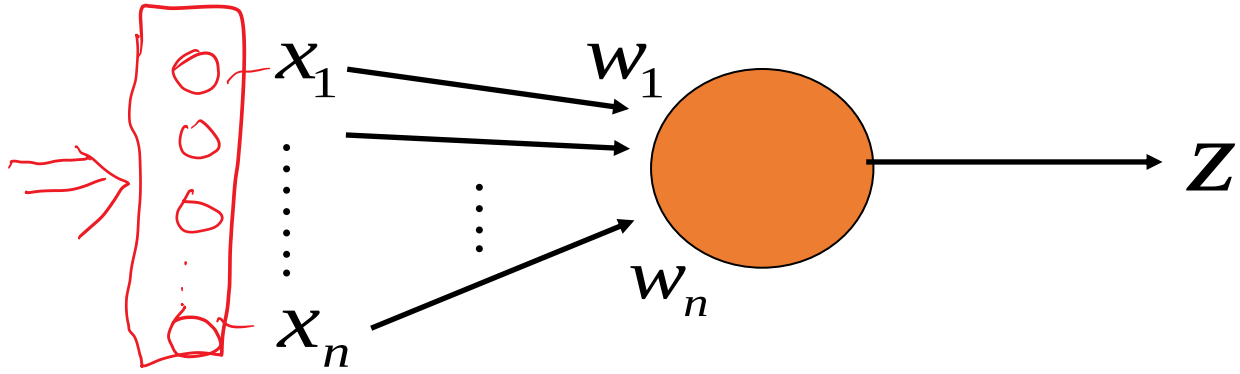


McCulloch-Pitts ニューロン （0，1の2値）

多層、フィードバック、側抑制

パーセプトロンの収束定理・問題点

H.D.Block, Minsky その他



3層、出力層のみが学習

Random結合の中間層 universal approximator

Extreme learning machine
Liquid machine

Minsky-Papert 批判 1969
連結性、凸性、ベッチ数

アナログパーセプトロンと中間層の学習

オンライン勾配降下法

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

Amari, 1966, 1967; Tsypkin, 1966

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

学習 $\Delta\theta_t = -\eta_t \frac{\partial l(x, \theta_t)}{\partial \theta}$

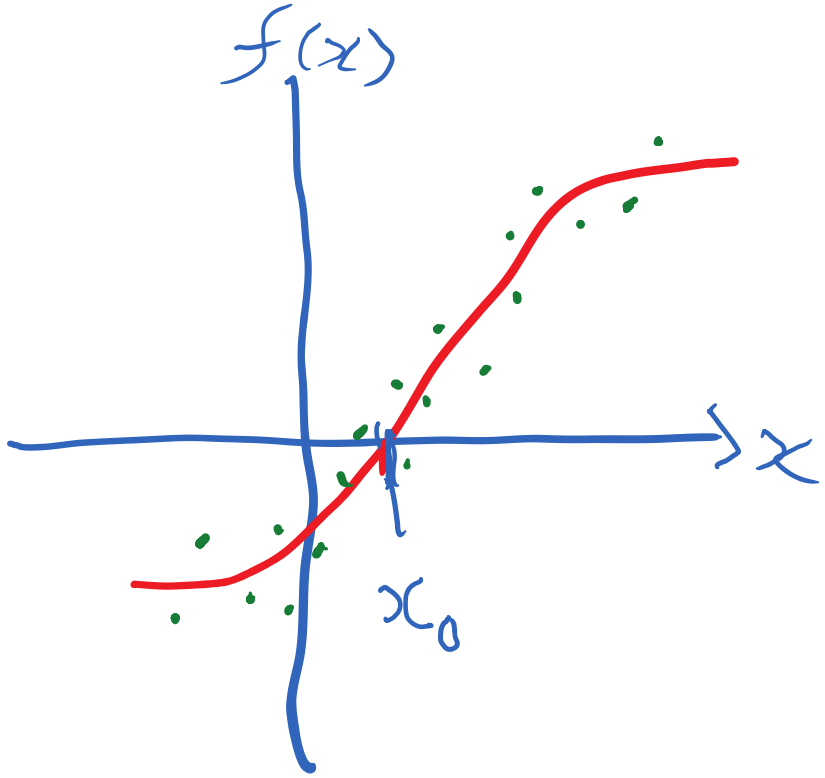
微分可能!

確率勾配降下法
Stochastic gradient descent

確率近似法　Robbins-Monro, 1951
Kiefer-Wolfwitz, 1952

Stochastic approximation

$$\Delta x_t = -\eta_t f(x_t)$$



$$\sum \eta_t > \infty$$

$$\sum \eta_t^2 < \infty$$

確率勾配降下学習法ーオンライン

Robbins-Monro確率近似法；（Kieferー Wolfwitz）

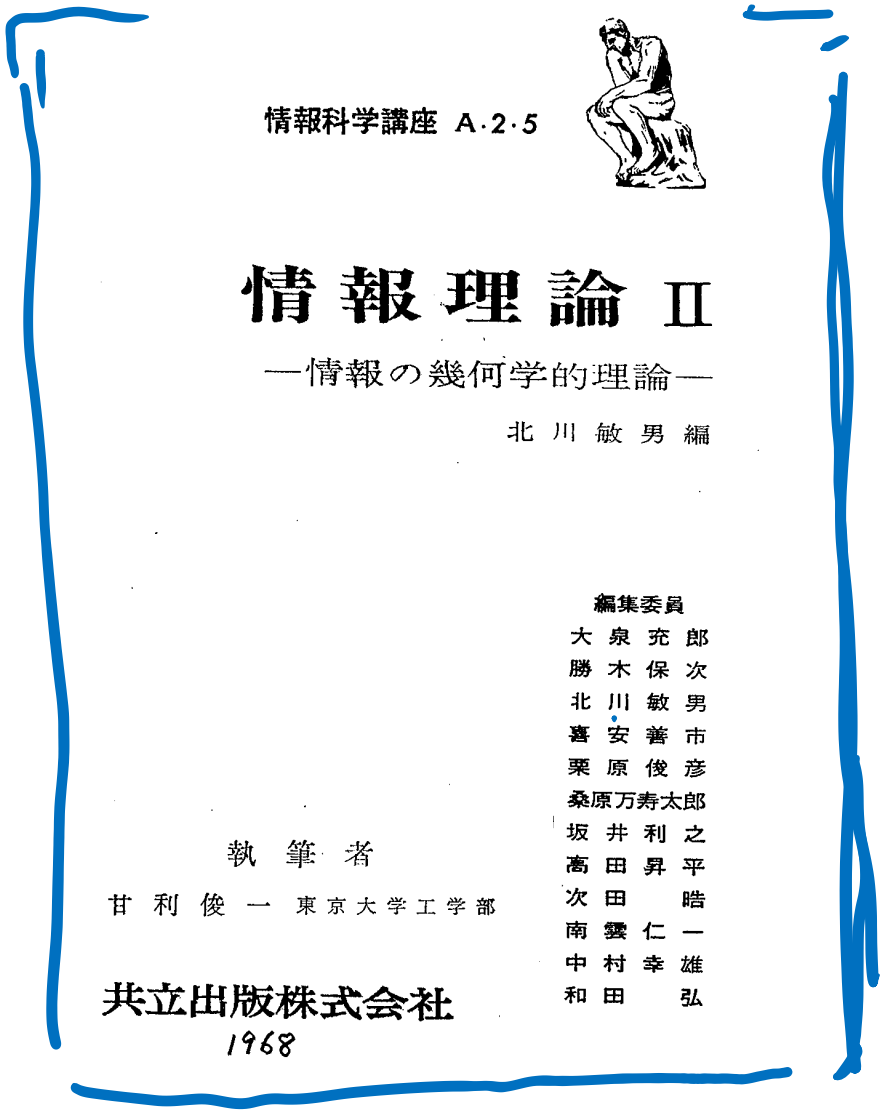
中間層の学習

機械学習　Amari, 1966、1967; Tsypkin, 1966;
Werbos, 1974

Error back-propagation --- PDP

Rumelhart, Hinton, Williams, 1986

最初のMLPの確率勾配降下学習法 (1966－1968)

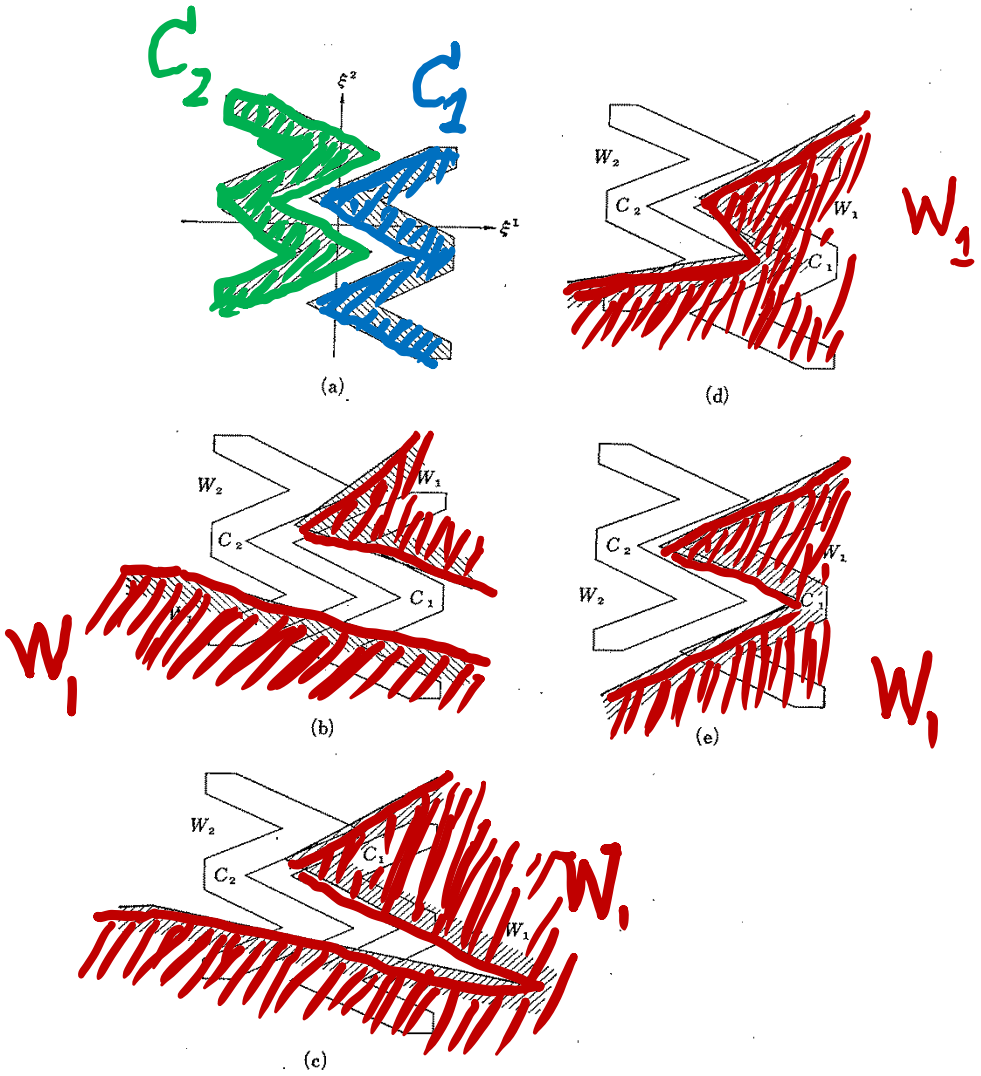


Information Theory II
--Geometrical Theory of Information

Shun-ichi Amari
University of Tokyo

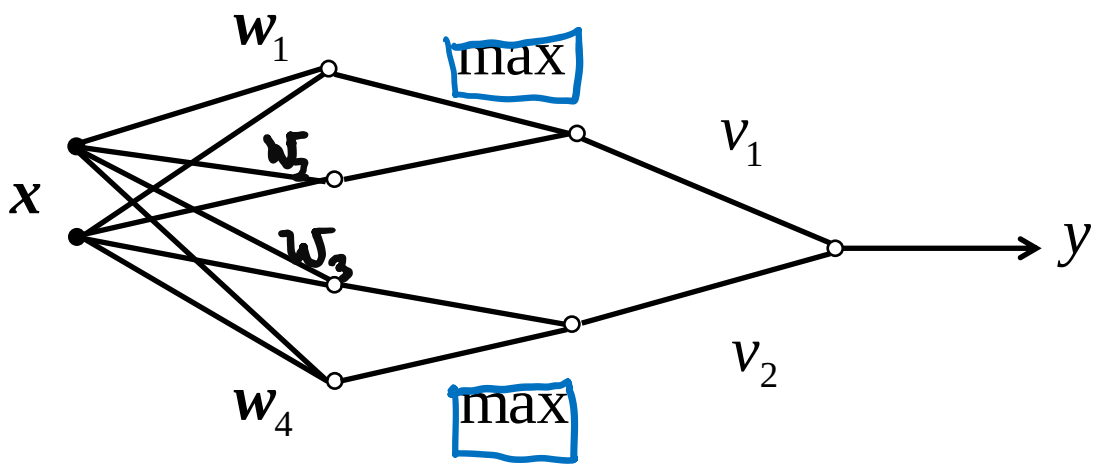
Kyouritu Press, Tokyo, 1968

線形分離不可能パターン分類



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

アナログニューロン、シグモイド関数



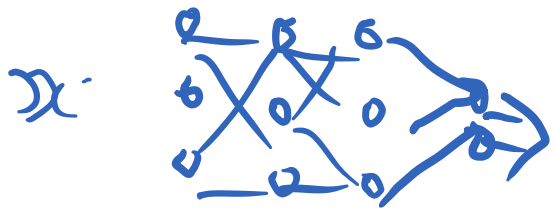
Error back-propagation --- PDP

Rumelhart, Hinton, Williams, 1986

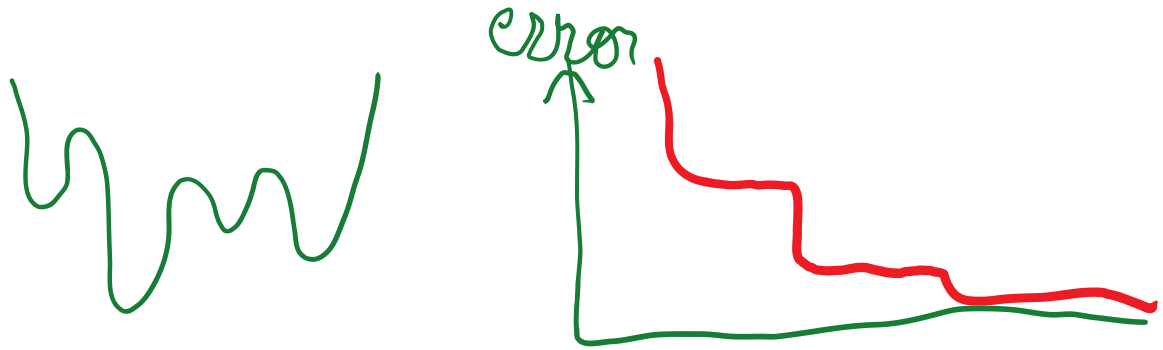
$$e = \frac{\partial f(x, \theta)}{\partial w}$$

$$\Delta w = -\eta e x$$

Amari, 1966; Tsypkin, 1966;
Werbos, 1974



Universal approximator; local minima; plateau



$$e^{l-1} = \frac{\partial f}{\partial w^l} e^l$$

深層学習の勝利　2010~

パターン認識
囲碁
文章生成・解析・翻訳

学習の理論的課題　1：汎化誤差

N:　例題数　　P：モデルのパラメータ数

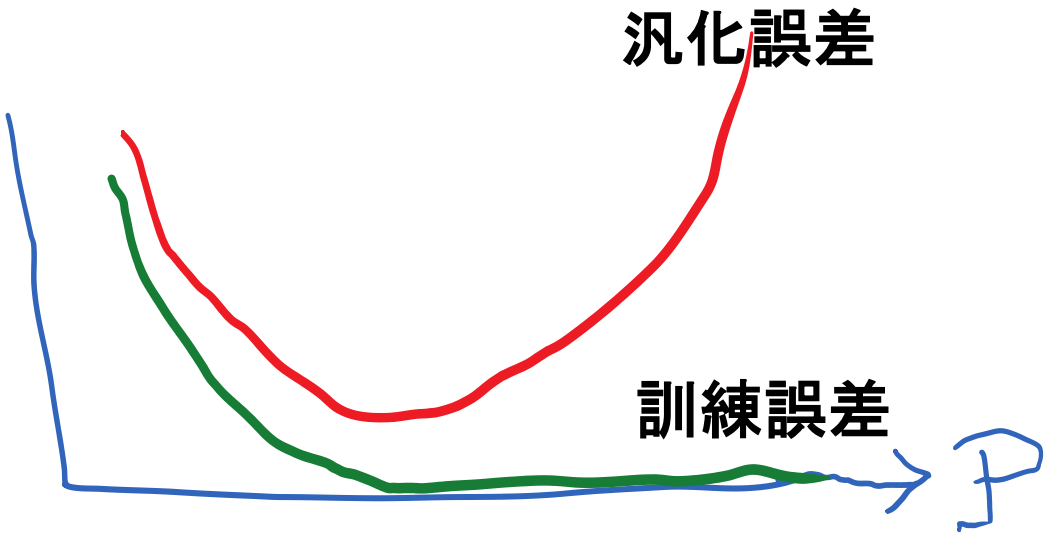
$$y = f(x,\theta) + \varepsilon$$

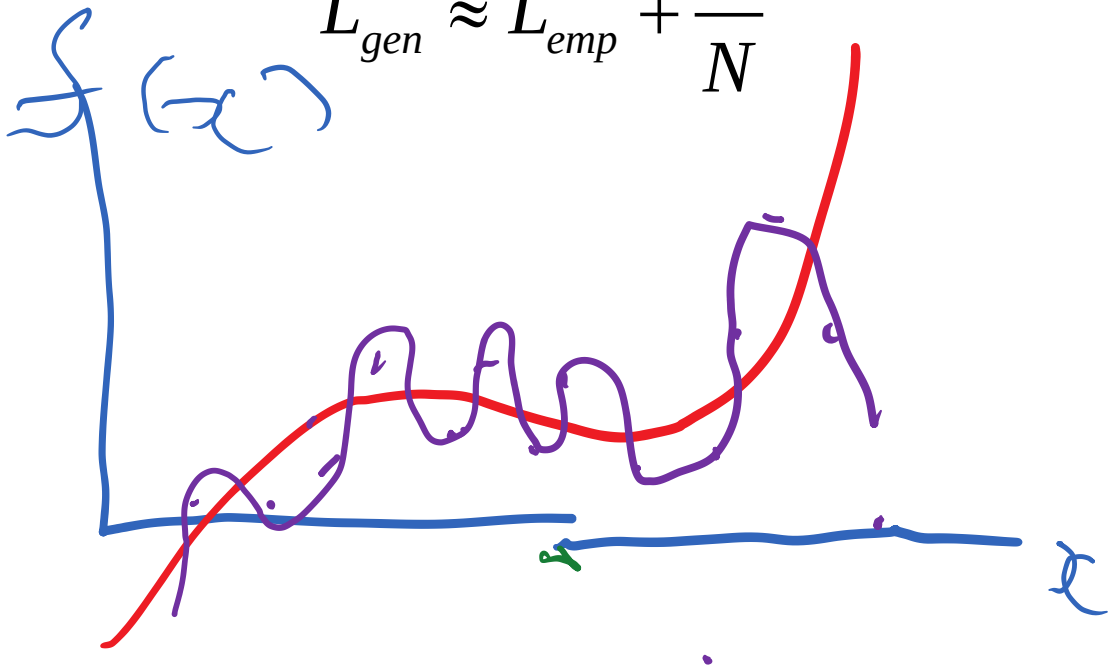
$$L_{emp} = \frac{1}{N}\sum |\, y_i - f(x_i,\theta)\,|^2$$

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

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

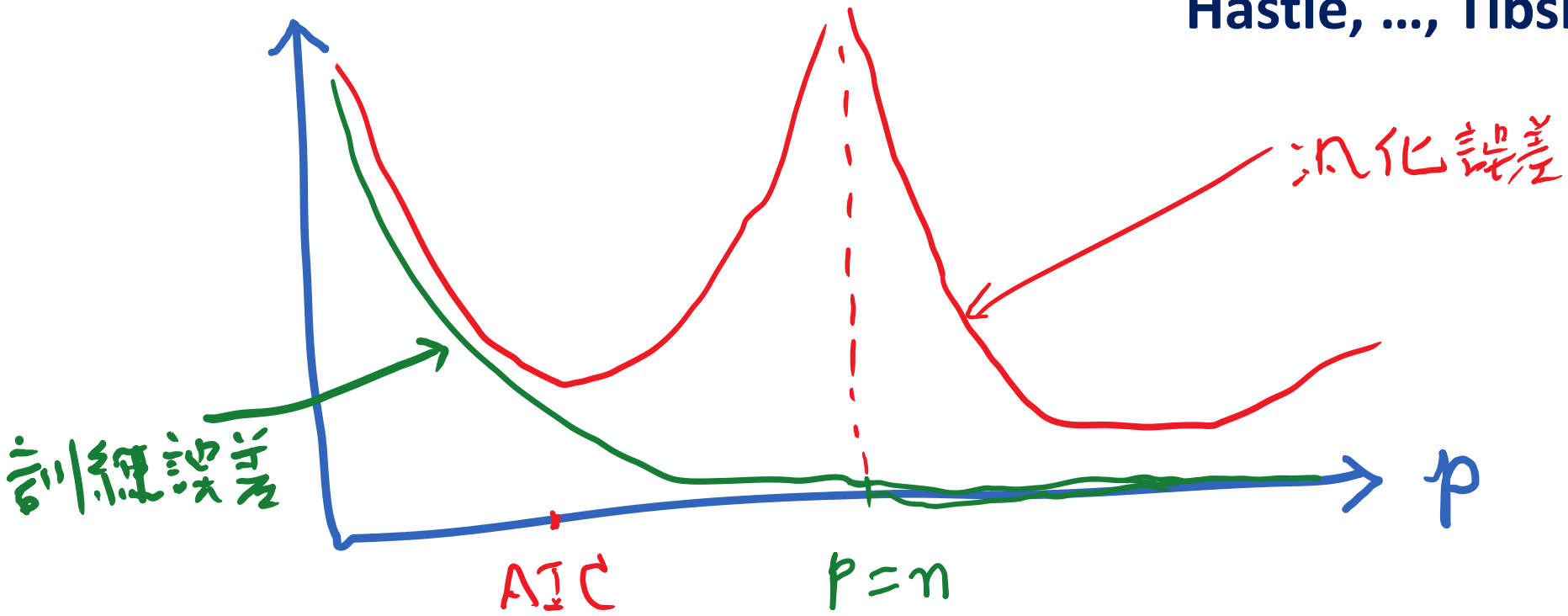
過学習





汎化誤差ー誤差曲線　double descent

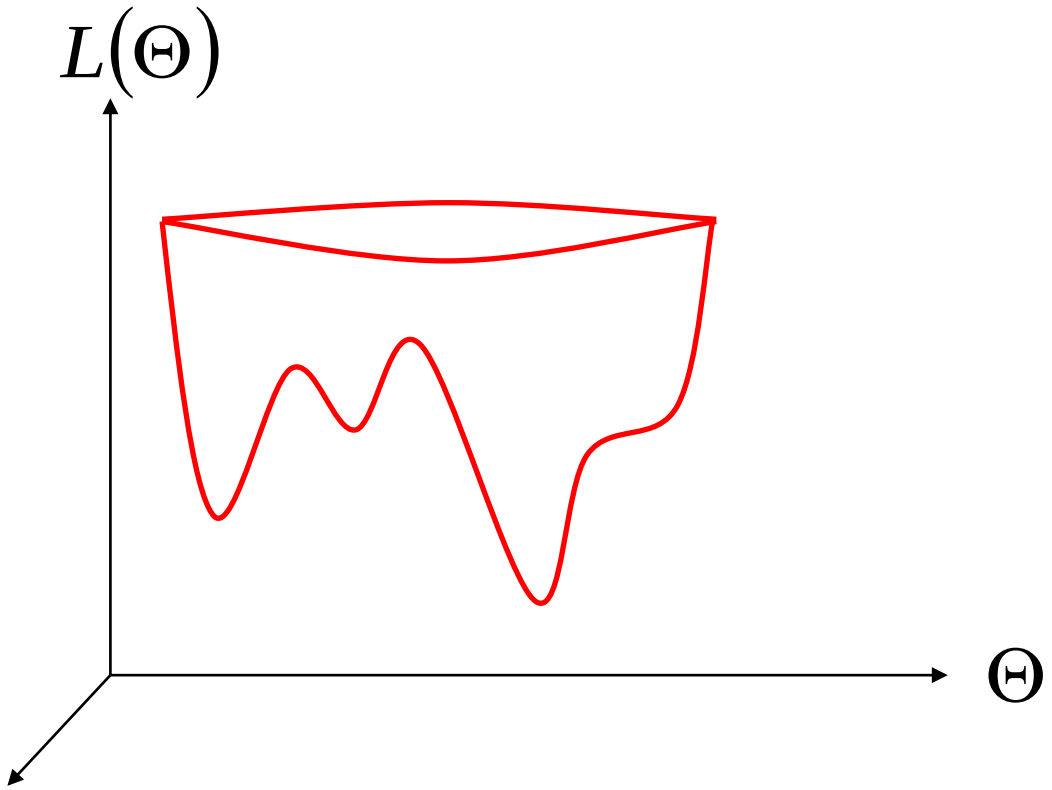
Belkin, Hsu, Xu; 2018
Hastie, …, Tibshrani; 2019



学習の理論的課題　2:　極小解と大域解

Simulated annealing

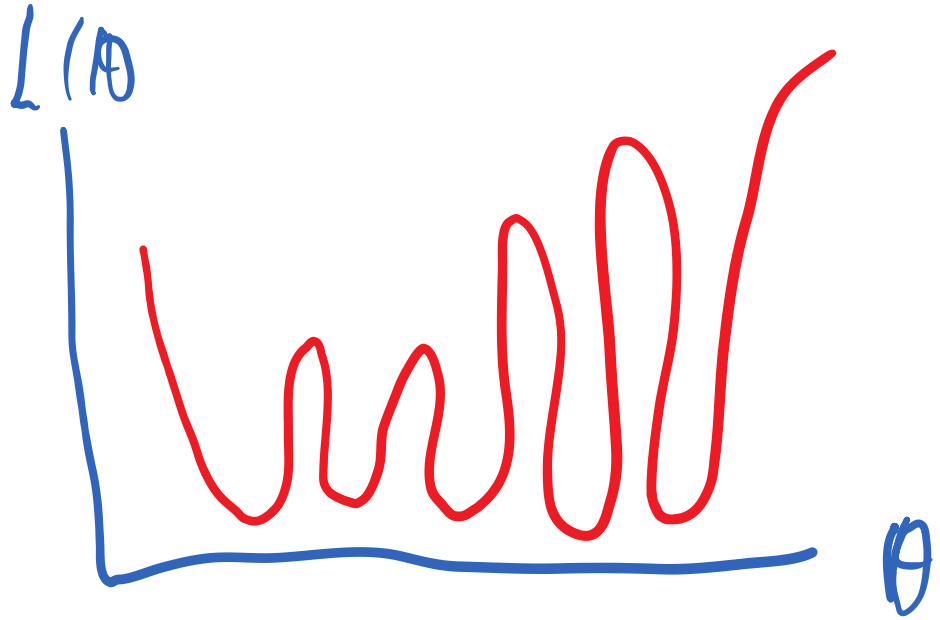
Quantum annealing



P大： P>>N

極小解 ≈ 大域解

Kawaguchi, 2019 ….



私の研究

1970年代　神経回路網の数理

統計神経力学　ランダム結合
連想記憶
自己組織化
神経場の力学と自己組織

情報幾何の始まり～ 1970年代後半　自分の可能性

情報幾何の発展

Chentsov, Efron, Dawid　（1970～1980）

推論の高次漸近理論

　Cox―統計の微分幾何（London会議）

情報幾何：応用分野の広がり

信号処理
学習
パターン認識

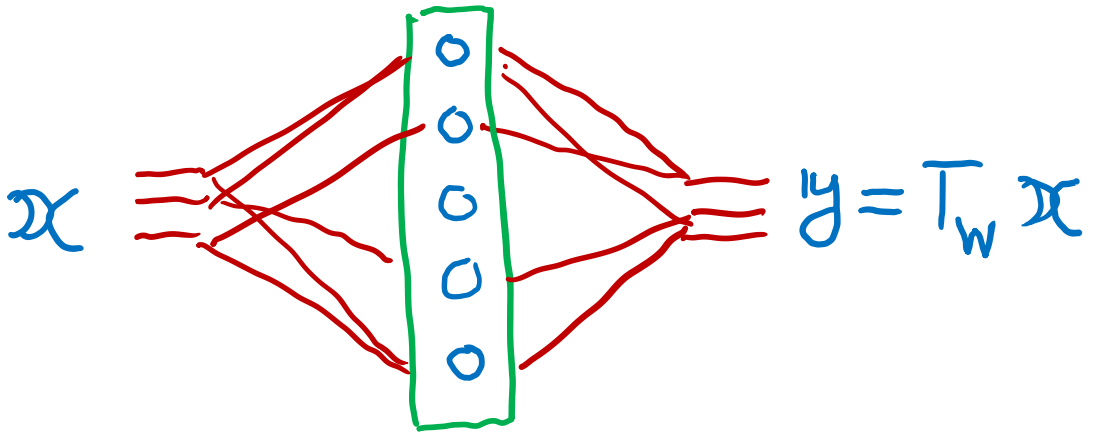
自然勾配学習ー人工知能

統計神経力学:ランダム結合

Rozonoer (1969）
Amari (1969、1971; 1974；…）
Amari et al (2013)
H. Sompolinski

Poole, …, Ganguli (2016)

S. Schoenholz, et al., 2017



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

巨視的振舞い

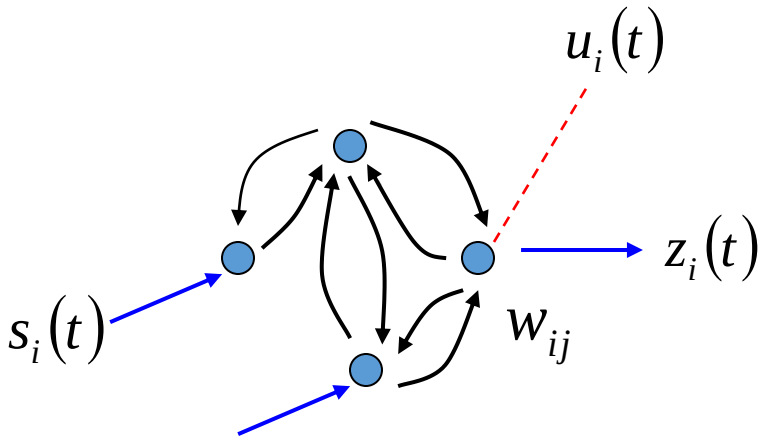
ほとんどすべての（典型的）回路に共通

神経集団の力学

$$\tau \dot{u}_i(t) = -u_i + \sum_j w_{ij} z_j(t) + s_i(t) - h$$

$$z_j(t) = f[u_i(t)]$$

$$\tau \dot{\mathbf{u}} = W\mathbf{f}(\mathbf{u}) + \mathbf{s} - \mathbf{h} - \mathbf{u}$$

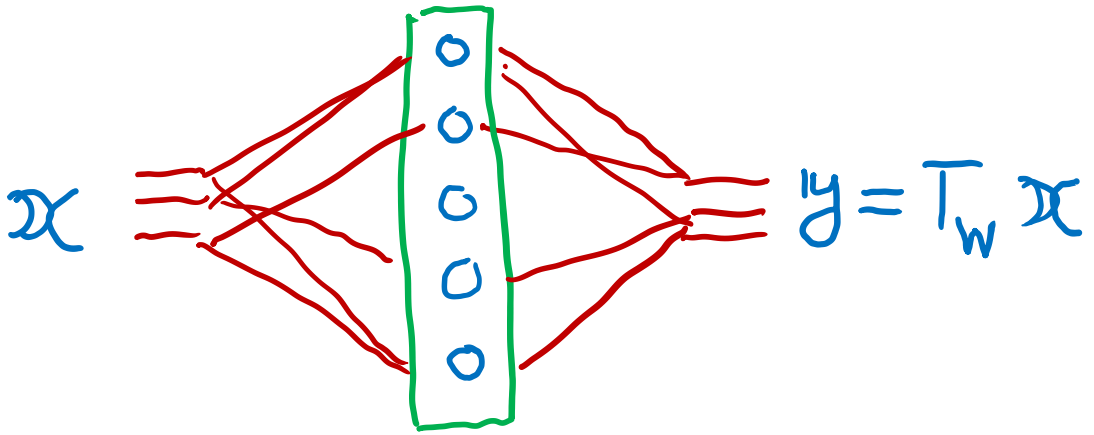


集団符号化:巨視的力学法則

W_{ij} : *random, iid*　巨視的状態:発火率 $Z = \frac{1}{n}\sum z_i$;　$U = \frac{1}{n}\sum u_i$

統計神経力学:ランダム結合

Rozonoer (1969)
Amari (1969、1971; 1974; …)
Amari et al (2013)
H. Sompolinski
Poole, …, Ganguli (2016)
S. Schoenholz, et al., 2017



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

巨視的振舞い

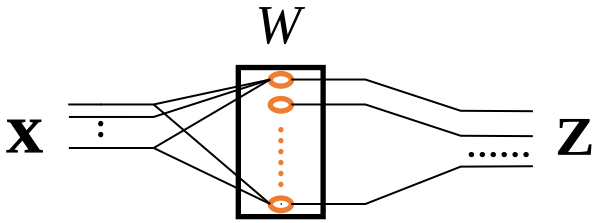
ほとんどすべての(典型的)回路に共通

Rozonoer　1969

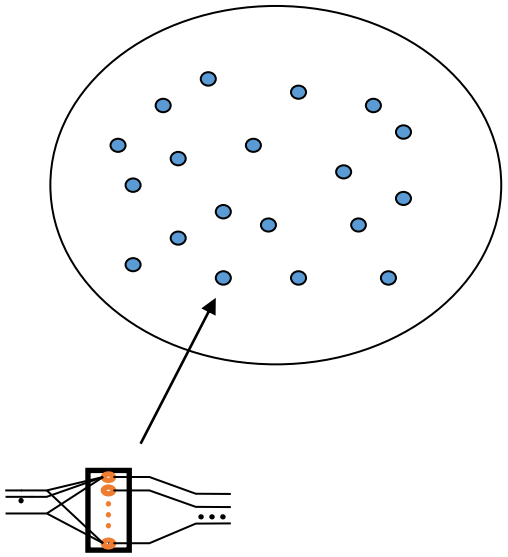
Boltzmann H 定理（エントロピー増大)

Amari(1969): 神経集団の力学、基礎付け

ランダム回路のアンサンブル

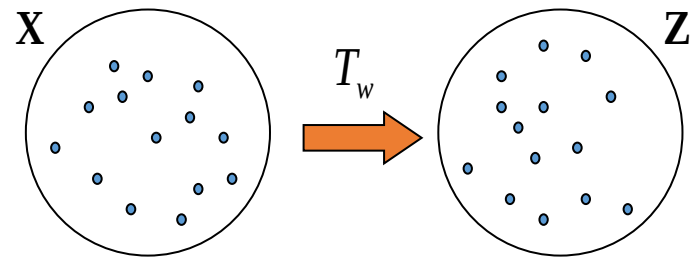


$$\Omega = \{w\}, \quad p(w)$$



微視的状態遷移

$$\mathbf{z} = T_w \mathbf{x}$$



巨視的状態

$$X = X(\mathbf{x}); \quad Z = Z(\mathbf{z})$$

$$X = \frac{1}{n}\sum x_i \quad , \quad Z = \frac{1}{n}\sum z_i$$

$$Z = Z(\mathbf{z}) = Z(T_w \mathbf{x})$$

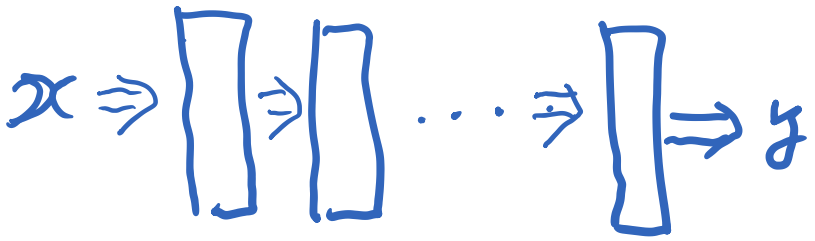
$$Z = F(X(\mathbf{x}))$$

$Z = F(X)$ ： 巨視的法則

深層学習の統計神経力学

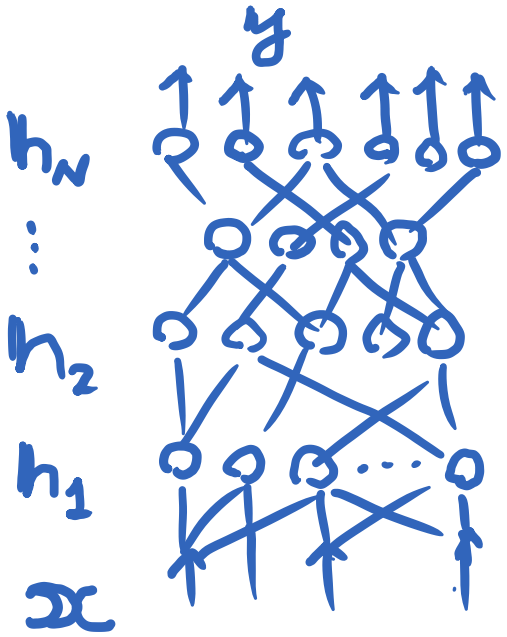
Poole, …, Ganguli (2016)――信号変換
S. Schoenholz, et al., 2017――誤差逆変換

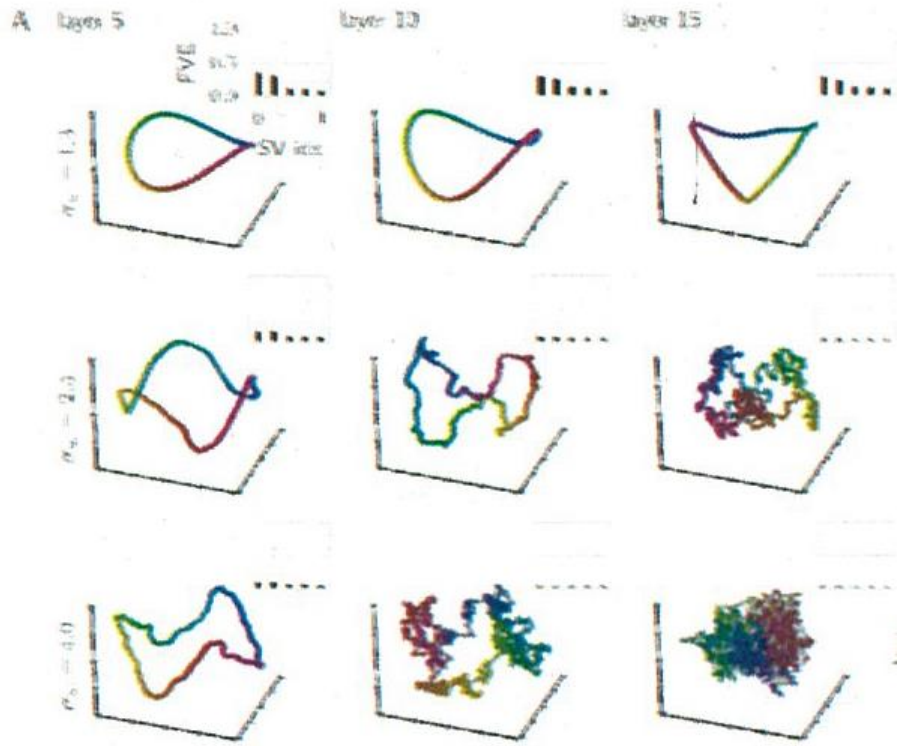
甘利、唐木田、赤穂



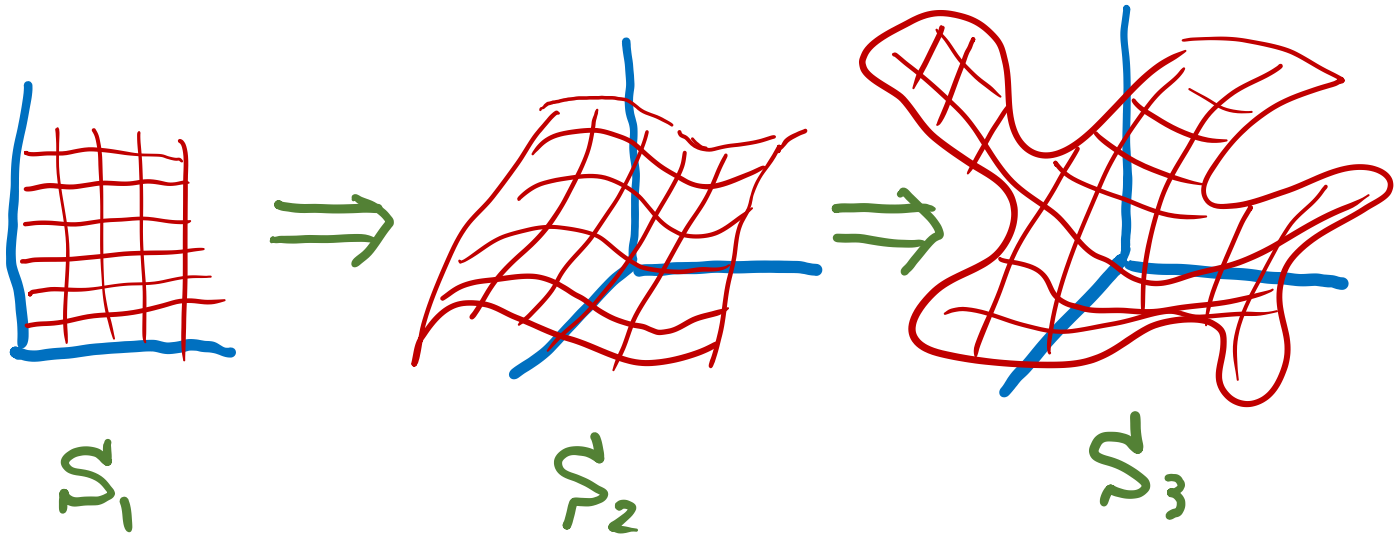
ランダム回路は万能！

Poole et al (2016)
Random deep neural networks





引き戻し計量（リーマン計量・距離）



$$ds^2 = \sum g^l{}_{ab}\, dx^a dx^b = \frac{1}{n_l} d\mathbf{x}^l \cdot d\mathbf{x}^l$$

$$g^l{}_{ab} = \mathbf{e}^l{}_a \cdot \mathbf{e}^l{}_b$$

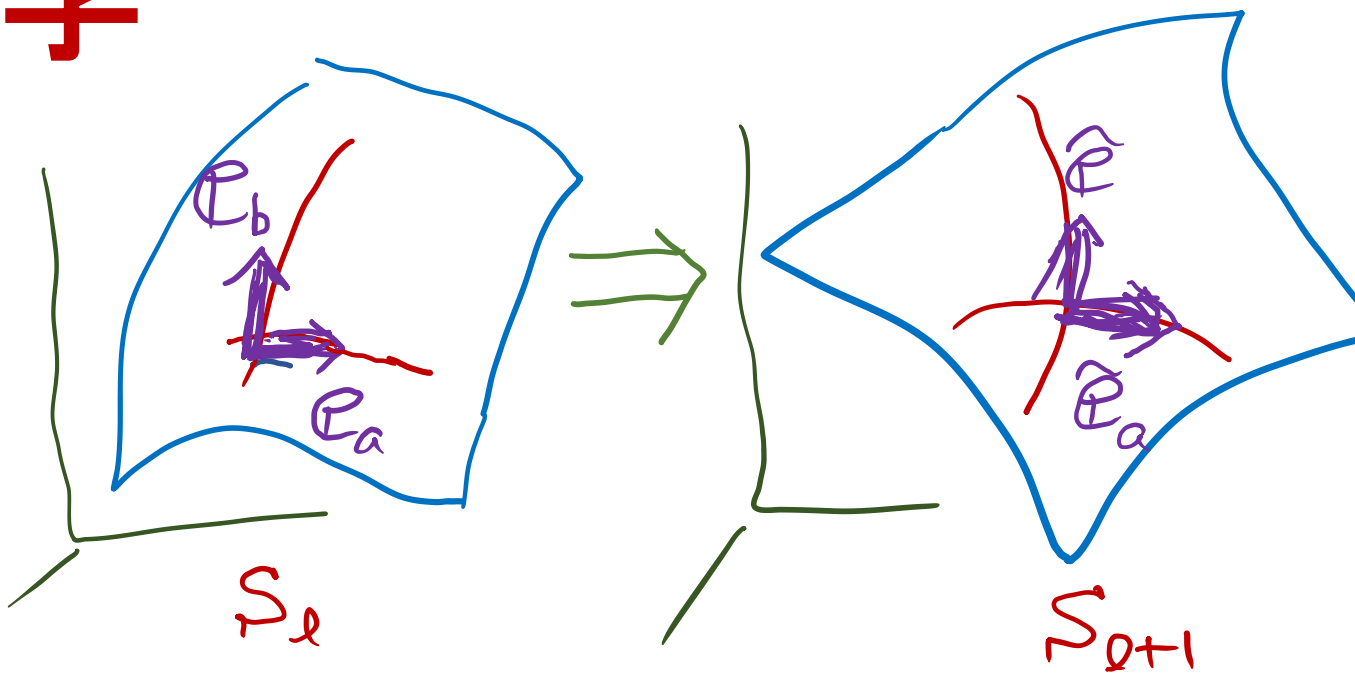
リーマン計量の力学

$$\tilde{y}_\alpha = \varphi(\sum w_{\alpha k} y_k + b_\alpha) = \varphi(u_\alpha)$$

$$d\tilde{y}_\alpha = \sum B_k^\alpha dy_k \quad \tilde{\mathbf{e}}_a = B\mathbf{e}_a$$

$$B = (B_k^\alpha) = (\varphi'(u_\alpha) w_k^\alpha)$$

$$\tilde{g}_{\alpha\beta} = \sum \delta_{\alpha\beta} B_k^\alpha B_j^\beta g_{kj} = <\tilde{\mathbf{e}}_\alpha, \tilde{\mathbf{e}}_\beta>$$



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

平均場近似不使用

$$\chi_1(A) = \int \sigma^2 \{\varphi'(\sqrt{A}v)\}^2 Dv = \frac{1}{2\pi} \frac{\sigma^2 A + \sigma_b^2}{\sqrt{1 + 2(\sigma^2 A + \sigma_b^2)}}$$

多層神経回路:パラメータ学習

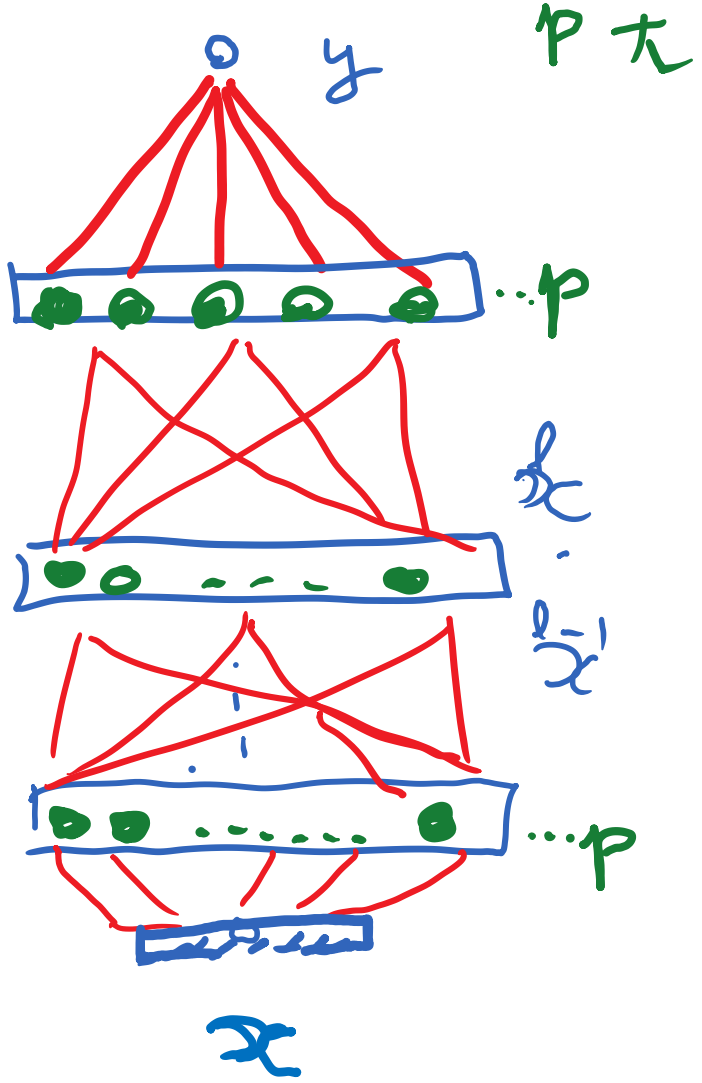
$$y = \sum^{L} v_i \phi(\mathbf{w}_i \cdot \mathbf{x})$$

$$\mathbf{x}^{l} = \phi(\mathbf{w}_i^{l} \cdot \mathbf{x}^{l-1})$$

$$\theta = (\mathbf{v}_i, w_{ij}) \square N(0, \frac{\sigma^2}{p})$$

$$y = f(\mathbf{x}, \theta)$$

$$\dot{\theta} = -\eta \partial_\theta l$$

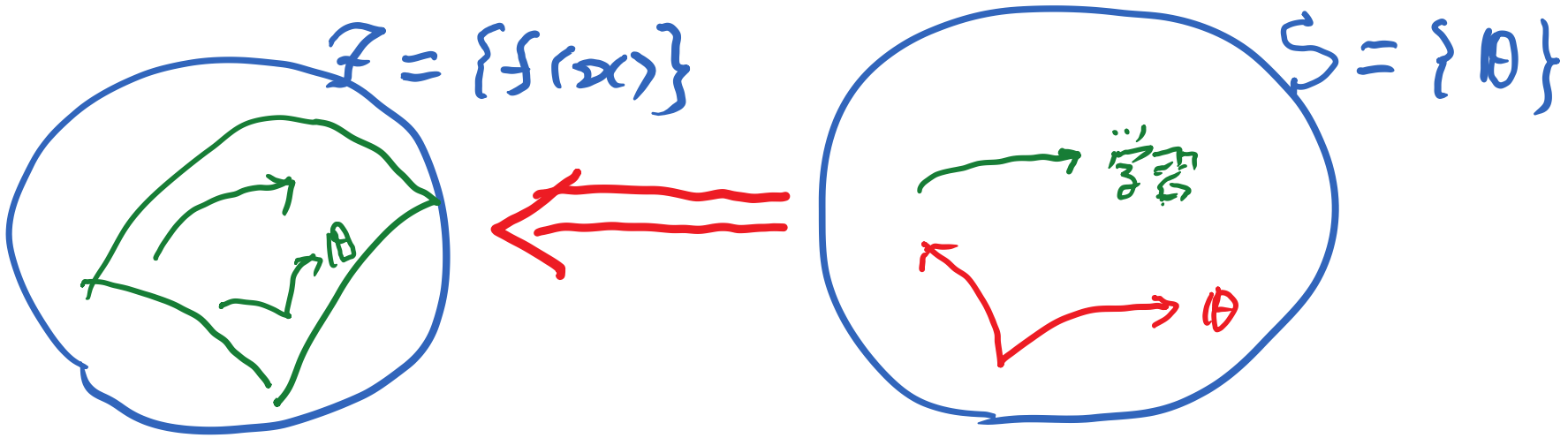


学習：関数空間とパラメータ空間ー最近の話題

NTK

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

$$f \Leftrightarrow \theta$$



Jacot et al; Neural tangent kernel

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

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

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

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

K; Gaussian kernel

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

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

ランダム回路のごく近傍に正解がある

ーーランダム深層回路

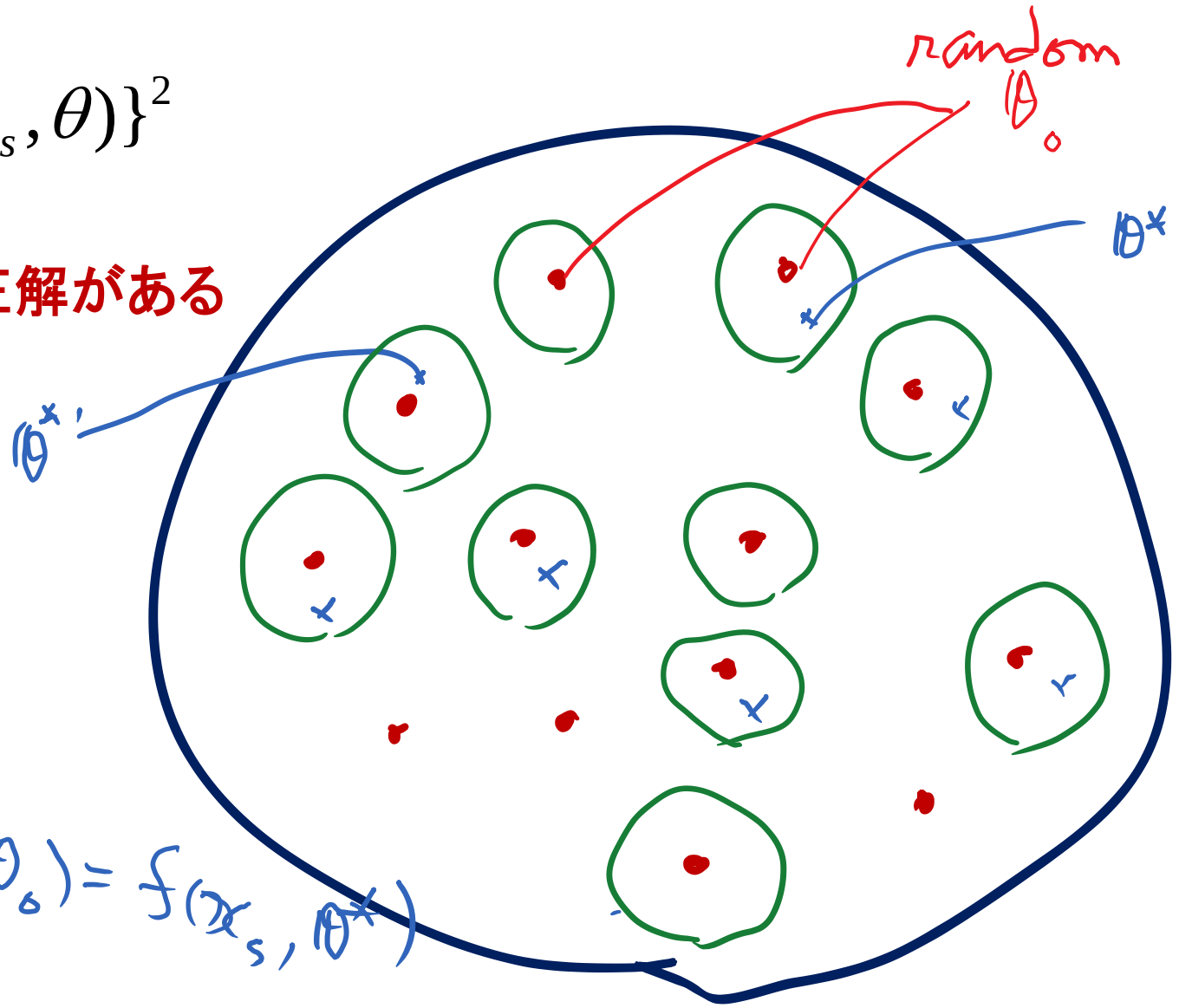
訓練データ　n 個　$D=\{(\mathbf{x}_1,y_1),\ldots,(\mathbf{x}_n,y_n)\}$

$$\theta^*=\arg\min\frac{1}{2}\sum_{s=1}^{n}\{y_s-f(\mathbf{x}_s,\theta)\}^2$$

定理　任意のランダム回路の近傍に正解がある

$$|\theta^*|^2,|\theta_0|^2\sim O(1)$$

$$|\theta_0-\theta^*|^2\square O(\frac{1}{p})$$

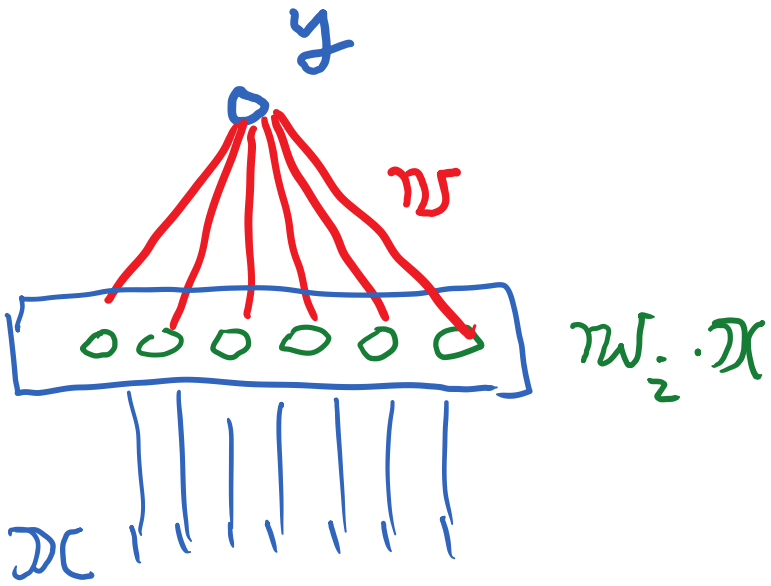


単純モデルによる線形理論

$$y = \sum v_i \phi(\mathbf{w}_i \cdot \mathbf{x})$$

$\mathbf{w}_i \sim$ random; fixed

$$v_i \square N(0, \frac{\sigma^2}{p}), \quad \sigma^2 = 1$$



補正

$$y_s^* = \sum (v_i + \Delta v_i)\phi(\mathbf{w}_i \cdot \mathbf{x}_s), \quad s = 1, \ldots, n$$

例題

$$e_s = y_s^* - \sum v_i \phi(\mathbf{w}_i \cdot \mathbf{x})$$

$$X_{si} = \phi(\mathbf{w}_i \cdot \mathbf{x}_s) = \frac{\partial f(\mathbf{x}_s, \theta)}{\partial \theta}$$

$$\mathbf{e} = X \Delta \mathbf{v}$$

$$\Delta \mathbf{v} = X^{\dagger} \mathbf{e}$$

$O(\frac{1}{p})$

$$\begin{bmatrix} \Delta v \end{bmatrix} = p \begin{bmatrix} X^{+} \end{bmatrix}^{n} \begin{bmatrix} e \end{bmatrix}$$

$$\begin{bmatrix} e \end{bmatrix} = n \overbrace{\begin{bmatrix} X_{si} \end{bmatrix}}^{p} \begin{bmatrix} v \end{bmatrix}$$

$$X^{\dagger} = X^T (XX^T)^{-1}$$

$O(\frac{1}{p})$

$$K = XX^T: \text{ tangent kernel } (K_{st} = K(\mathbf{x}_s, \mathbf{x}_t))$$

$O(p)$

$$\hat{f} = Ke = \partial_\theta f(x) \partial_\theta f(x') e$$

$$\sum_i \frac{\partial f(x_s)}{\partial_i} \frac{\partial f(x_t)}{\partial_i}$$

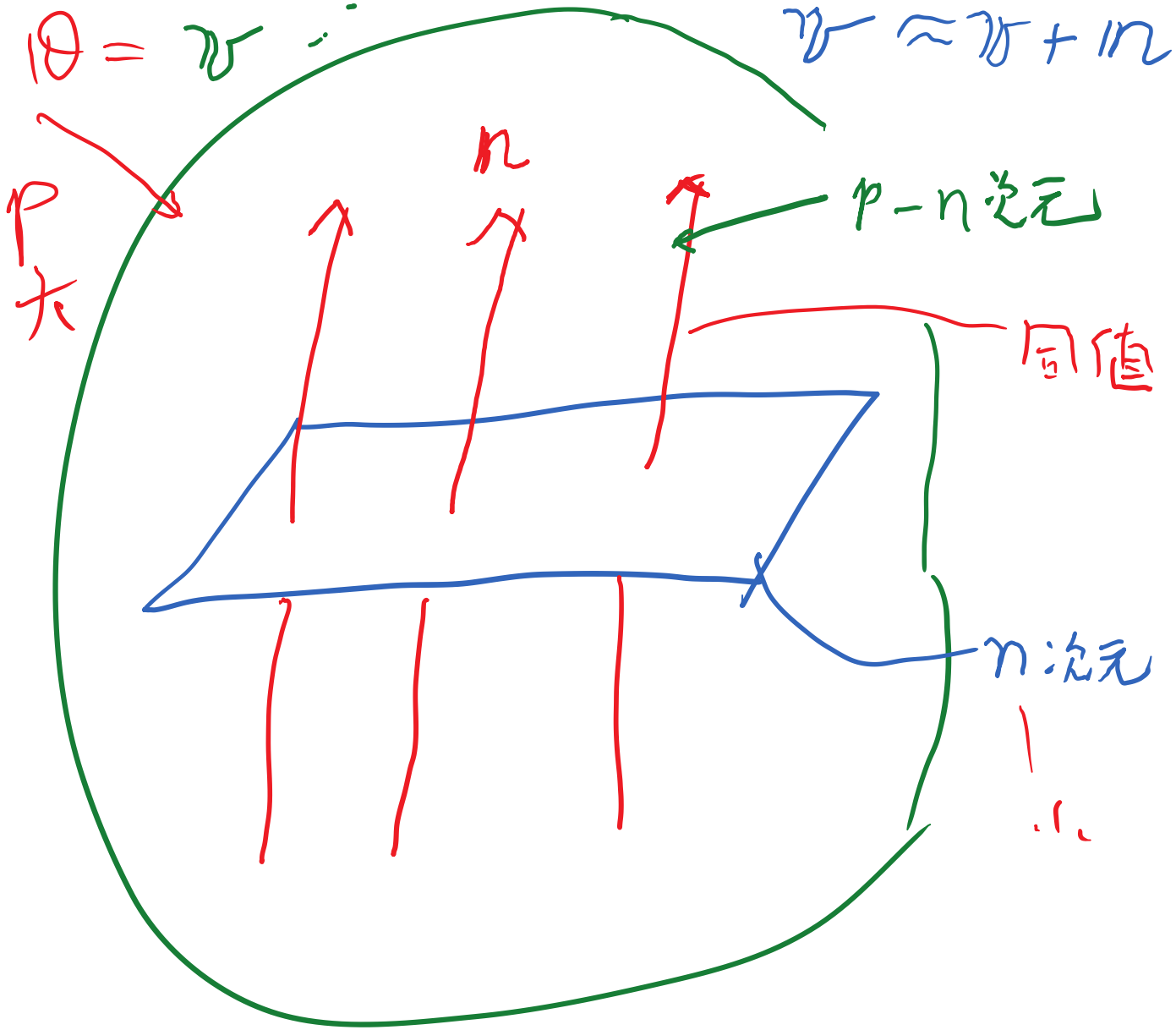
$$K^{-1}\mathbf{e} \sim O(\frac{1}{p}); \quad \Delta\mathbf{v} \sim O(\frac{1}{p})$$

零空間 N

$$e = X\Delta v$$

$$N = \{\mathbf{n} \mid \mathbf{Xn} = 0\}$$

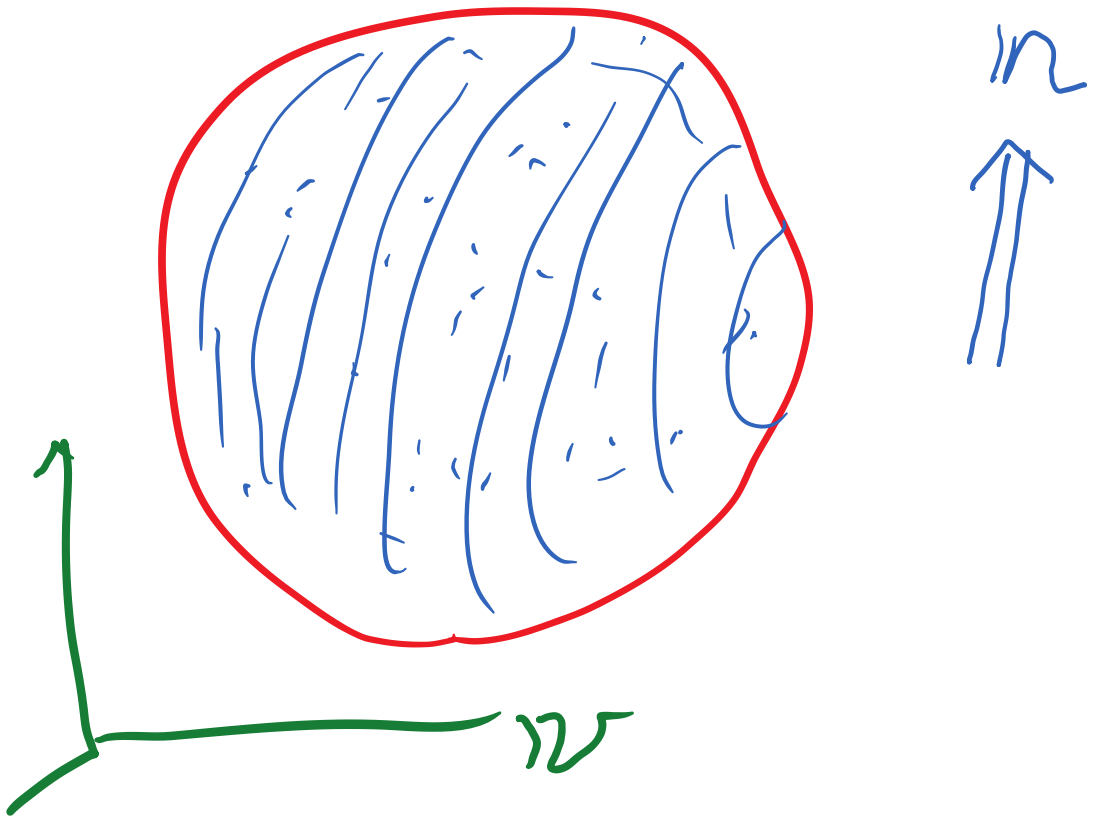
$$\Delta\mathbf{v} = X^{\dagger}\mathbf{e} + \mathbf{n}$$



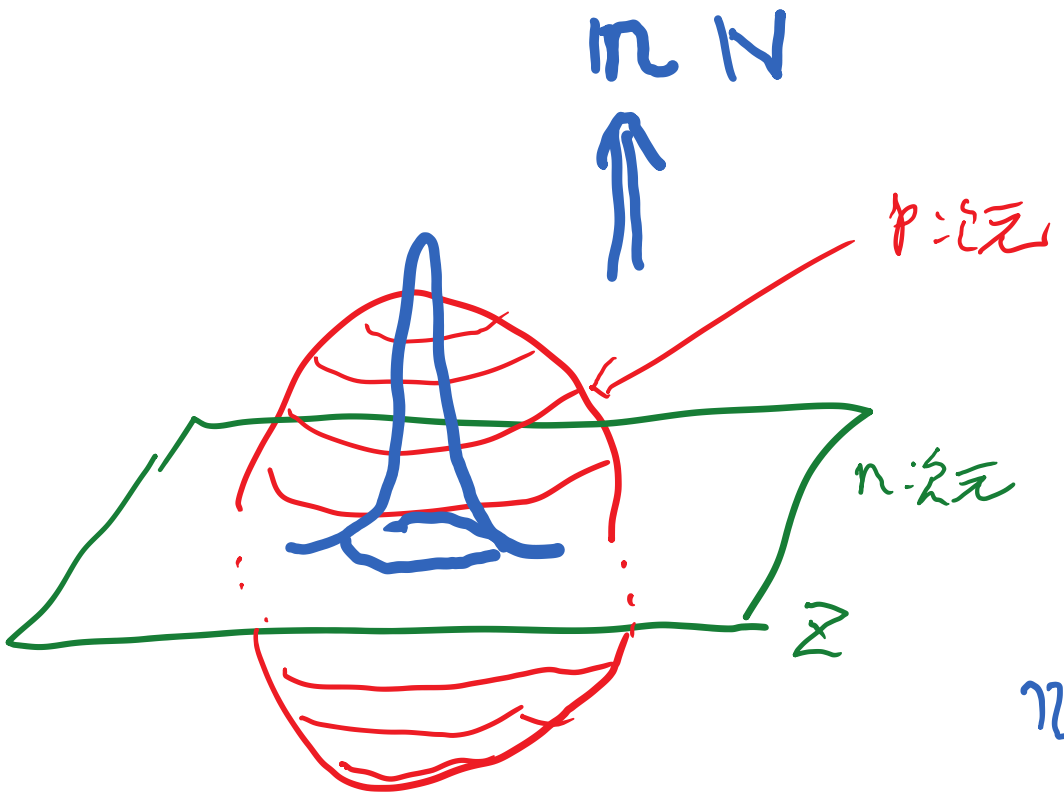
ランダム初期値 v の分布

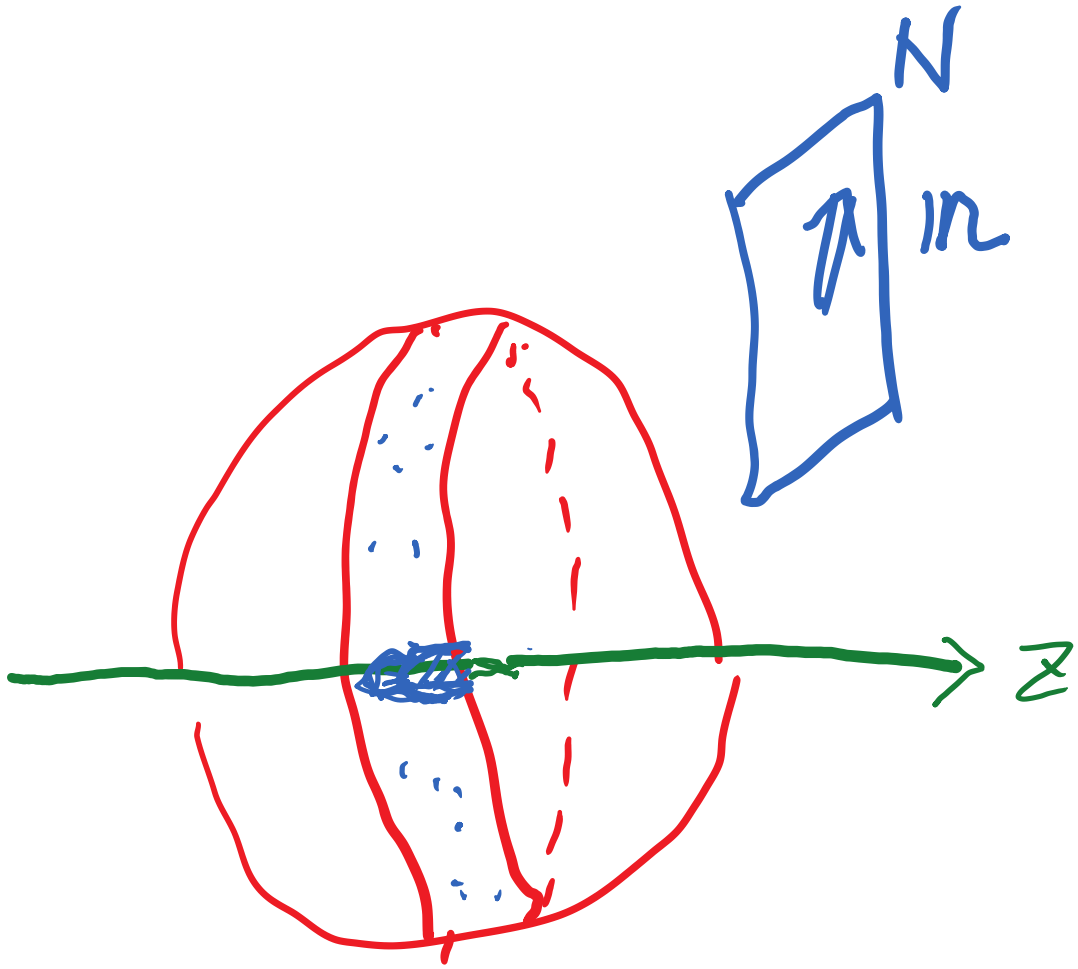
$$v_i \sim N(0, \frac{1}{p})$$

$$\sum v_i^{\,2} = 1$$



$$q(z) = c \exp\{-\frac{z^2}{2p}\}$$





$w + \Delta w$

一般の深層回路

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

$$\theta = (v_i, \mathbf{w}_i)$$

$$\Delta f = X \Delta\theta; \quad X = [\phi(\mathbf{w}_i \cdot \overset{L}{\mathbf{x}}) \quad v_i \phi'(\overset{L}{\mathbf{w}_i} \cdot \overset{L-1}{\mathbf{x}}) \overset{L-1}{\mathbf{x}} \quad \square\square\square \quad]$$

$$\Delta\theta = X^{\dagger}\mathbf{e} \quad X^{\dagger} = X^{T}(XX^{T})^{-1}$$

$$K = XX^{T}$$

$$\Delta\theta = O(\frac{1}{p}) = O(\frac{n}{Lp})$$

一般の深層回路

$$X = [\overset{L+1}{X}\ \overset{L}{X} \cdots \overset{l}{X} \cdots \overset{1}{X}]$$

$$\Delta\theta = X^{\dagger}\mathbf{e} \quad X^{\dagger} = X^{T}(XX^{T})^{-1}$$

$$K = XX^{T} = [\overset{L+1}{K}\ \overset{L}{K}\ \overset{L-1}{K} \cdots \overset{l}{K} \cdots \overset{1}{K}]$$

$$\Delta\theta = O(\frac{n}{pL})$$

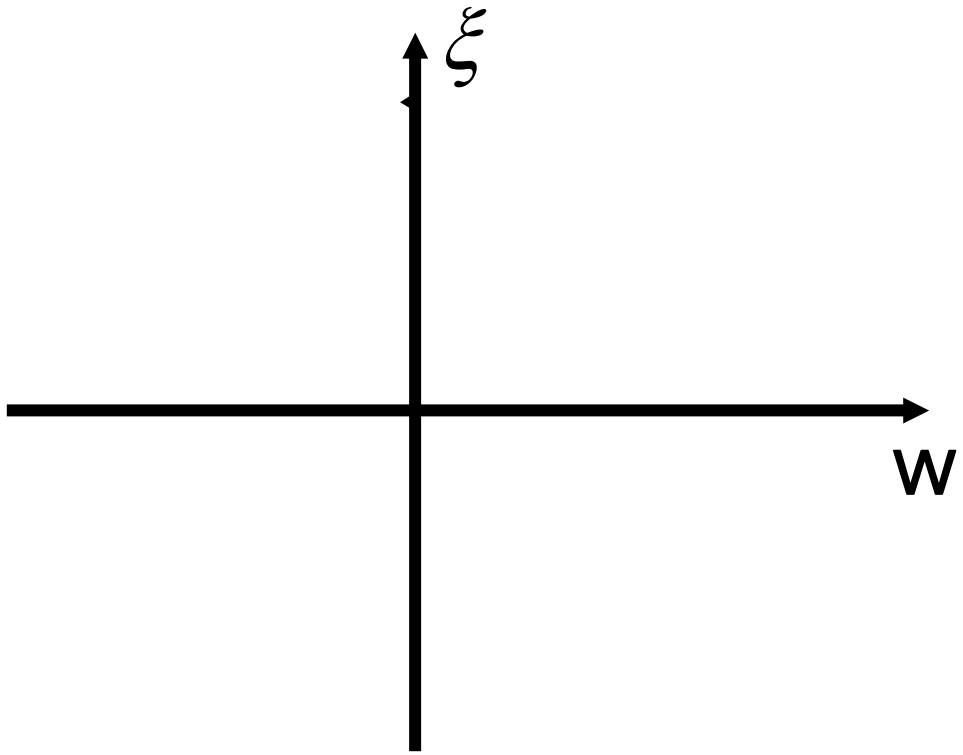
S. Amari, Any target functions exists in small neighborhood of random networks Neural computation, 2020

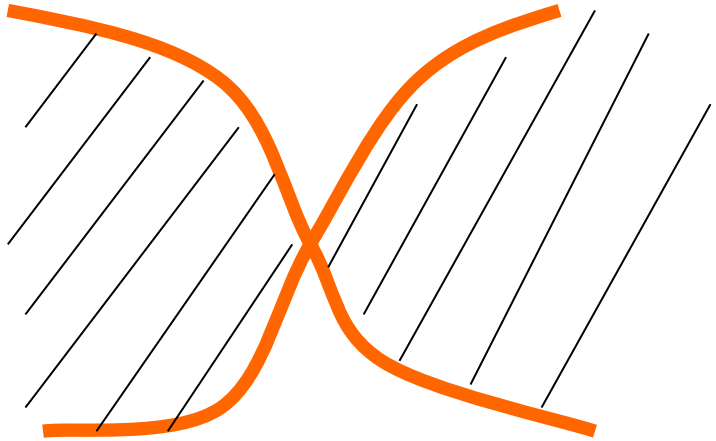
特異モデルの例

Milnor attractor

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

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



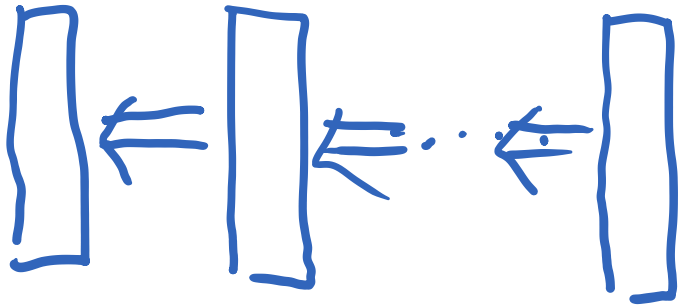


Fisher 情報行列と逆向き情報伝播

$$\frac{\partial l(x,W)}{\partial W^{l}} = e^{l}\, x^{l-1}$$

$$e^{l} = J^{l+1}\, e^{l+1};\ J = \frac{\partial f^{l+1}}{\partial W^{l}}$$

$$\Delta W^{l} = -\eta\, e^{l}\, x^{l-1}$$



error $e^{L} = y - x^{L}(W)$

$$l(x,W) = \frac{1}{2}\left|y - \varphi(x;W)\right|^2 = \left|e(x,y)\right|^2 = -\log p(x,W)$$

Fisher情報行列

ーー自然勾配学習法、ランダム深層回路

確率分布族

$$S = \{ p(x, \theta)$$

Fisher情報行列

$$g_{ij} = E[\partial_i \log p(x, \theta) \partial_i \log p(x, \theta)]$$

Cramer-Raoの定理

$$E[(\hat{\theta} - \theta)^2] \geq (g_{ij})^{-1}$$

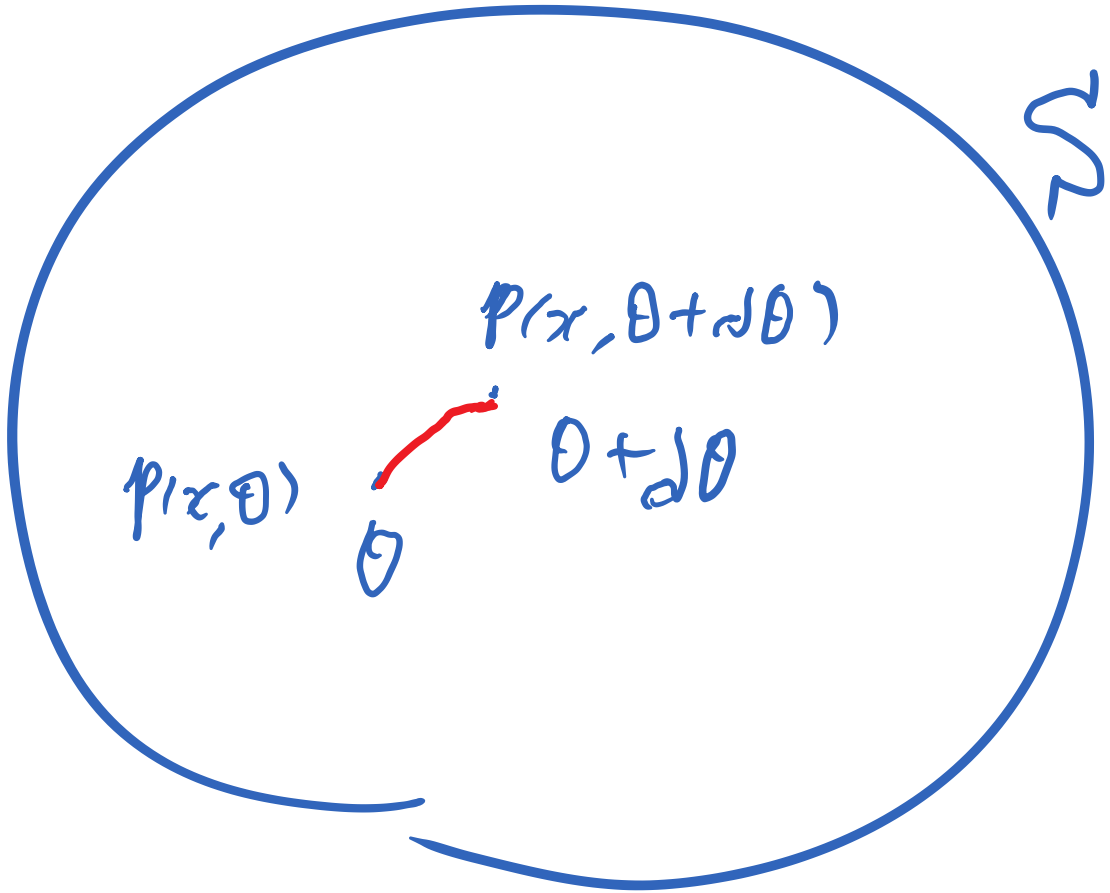
情報幾何

$$ds^2 = \sum g_{ij}(\theta) d\theta^i d\theta^j$$

確率分布族の空間

Riemann空間

微小距離



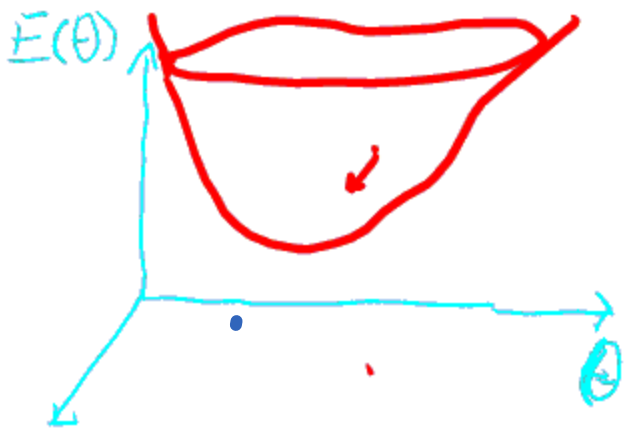
Natural Gradient

min $E(\theta)$

$$\max \quad |dE| = |E(\theta + d\theta) - E(\theta)|$$

$$|d\theta|^2 = \varepsilon = d\theta^T G \, d\theta$$

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



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

Information Geometry of MLP

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

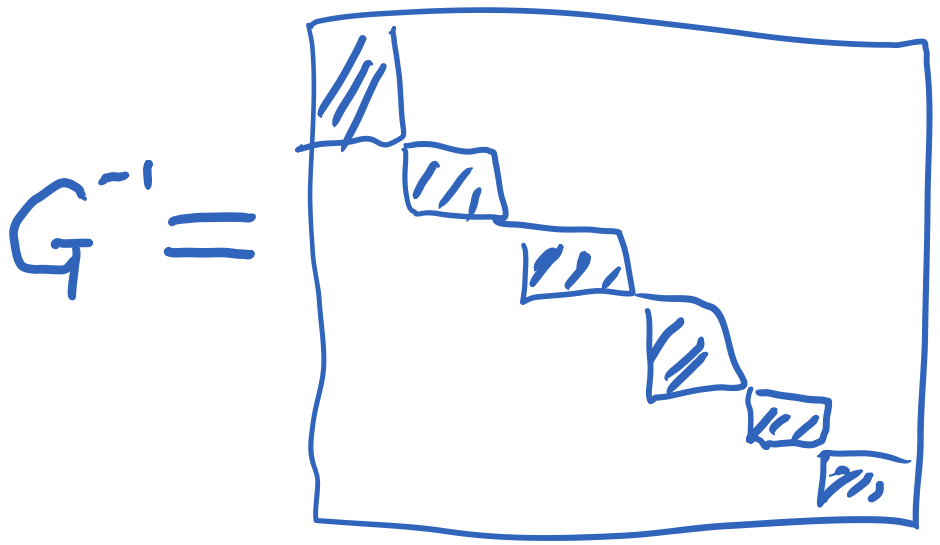
$$\Delta\boldsymbol{\theta} = -\eta G^{-1}(\boldsymbol{\theta})\frac{\partial l}{\partial \boldsymbol{\theta}}$$

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

Yann Ollivier

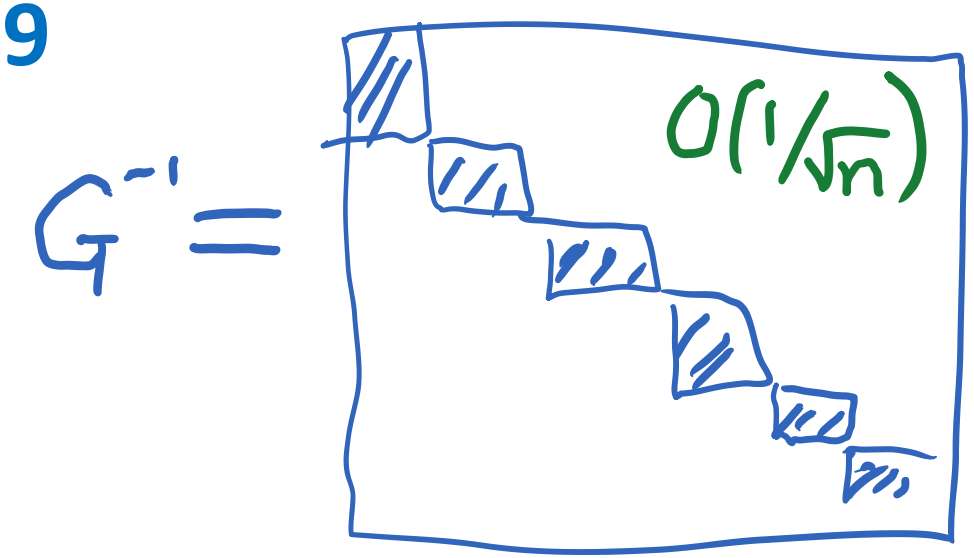
Block-wise diagonal
Unitwise diagonal
Bias+diagonal

KFAC
Arata



ランダム結合の神経回路

Amari, Karakida, AISTAT 2019



唐木田（産総研）の考え

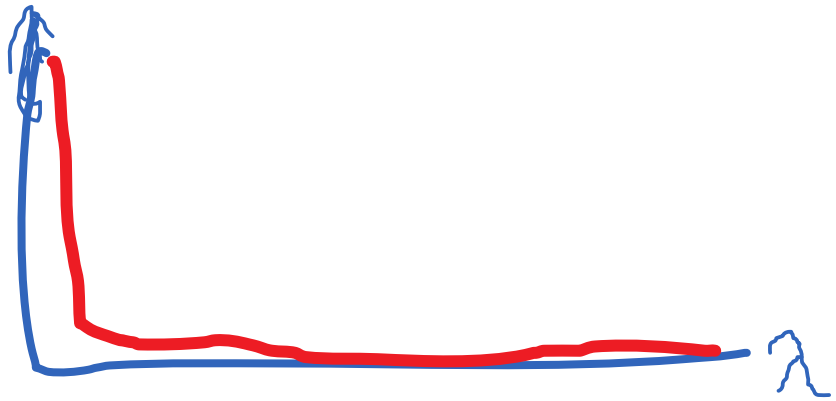
Schoenholz et al

$$F = E[\frac{d}{d\theta}\log p(y \mid x)\frac{d}{d\theta}\log p(y \mid x)]$$

$$p = \frac{1}{2}\exp\{y - f(x;\theta)\}^2$$

Fisher情報行列とバックプロップ

Fの固有値　ほとんど0と大きな値



誤差関数のHessian

$$\theta_{t+1} = \theta_t - \eta(y - f)\nabla f$$

error

Empirical Fisher 情報量

$$X(\chi,\theta)=[\nabla_\theta f(x_1,\theta),\nabla_\theta f(x_1,\theta),...,\nabla_\theta f(x_n,\theta)]$$

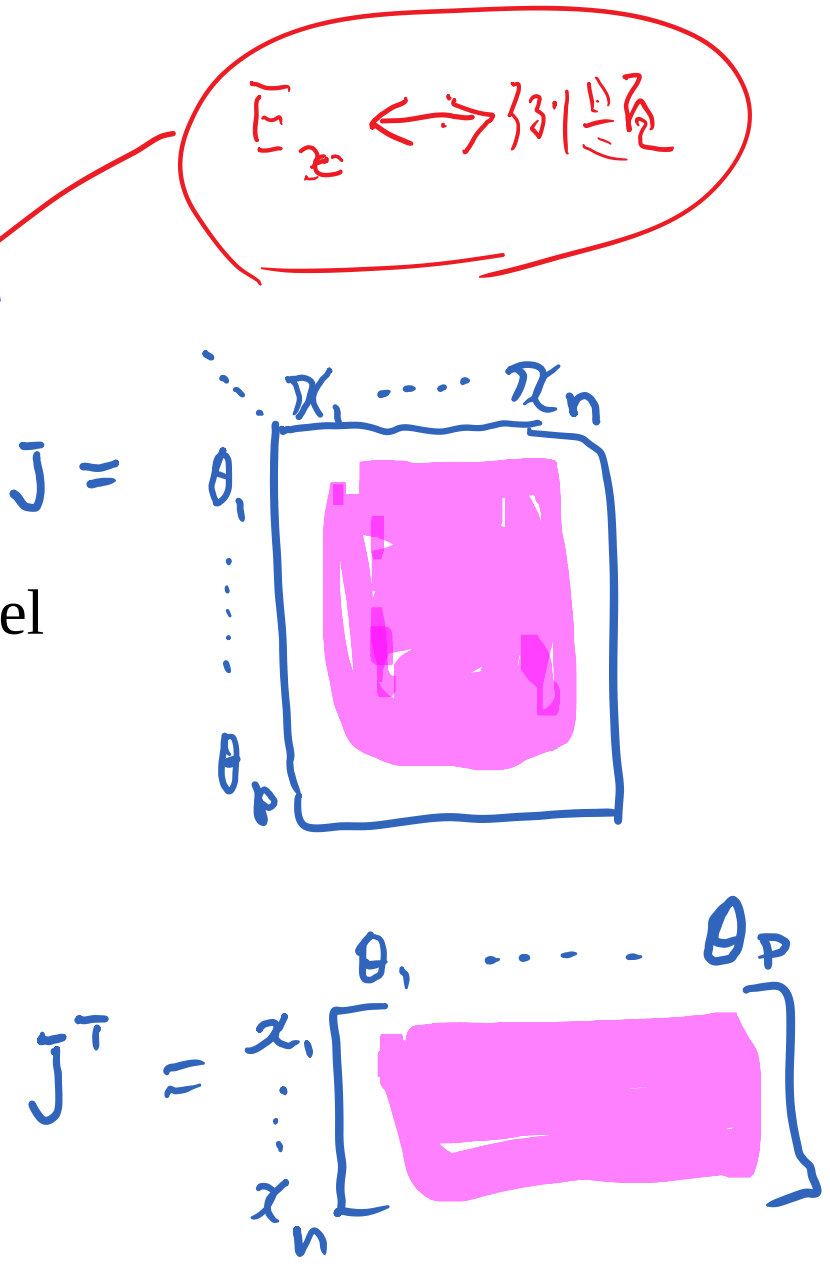
$$F(\theta)=\frac{1}{n}X^TX=\frac{1}{n}\sum\nabla_\theta f(x_s,\theta)\nabla_\theta f(x_s,\theta):\text{Fisher}$$

$$K(x,x')=XX^T=\sum{}_\theta\nabla_\theta f(x_s,\theta)\nabla_\theta f(x_t,\theta):\ \text{NT kernel}$$

$$F^\dagger=nX^T(XX^T)^{-1}$$

$$\Delta\theta=-\eta F^\dagger\nabla l=-\eta X^TK^{-1}e$$

G. Zhang, J. Martens and R. Grosse, 2019
T. Cai et al. 2019



Extended Adam

mini batch

$$l = \frac{1}{2} \sum K^{-1}{}_{st} e_s e_t$$

$$\Delta\theta = -\eta \partial_\theta l$$

Empirical natural gradient

深層学習：問題点と将来

大量のデータ、計算力

実験式の生成

入力を基に正解

現象の予測

日蝕の予測

ケプラーの法則、ニュートン力学

原理の創出・理解ー人間

深層学習はブラックボックスか？

だから危険か？

仕組みは分かっている、式は再現可能

現象の原理は示さない

数理脳科学：脳の基本原理の探求

単純な基本モデル用いる：数理的探索（現実とは違う）

→　計算論的神経科学
　　（脳はいかにこの原理を実現したか）

→　AI：技術による原理の実現　（脳とは違う）

脳は基本原理をどう実現したか

進化によるランダムサーチ
使える材料の制約
歴史的な制約

ごたごたの設計の中で精妙な実現：超複雑

人工知能は何をどう実現するか？

人工知能は脳に何を学ぶのか： 心
意識と無意識のダイナミックス

記号 --- 興奮パターン
論理的推論 --- 並列ダイナミックス

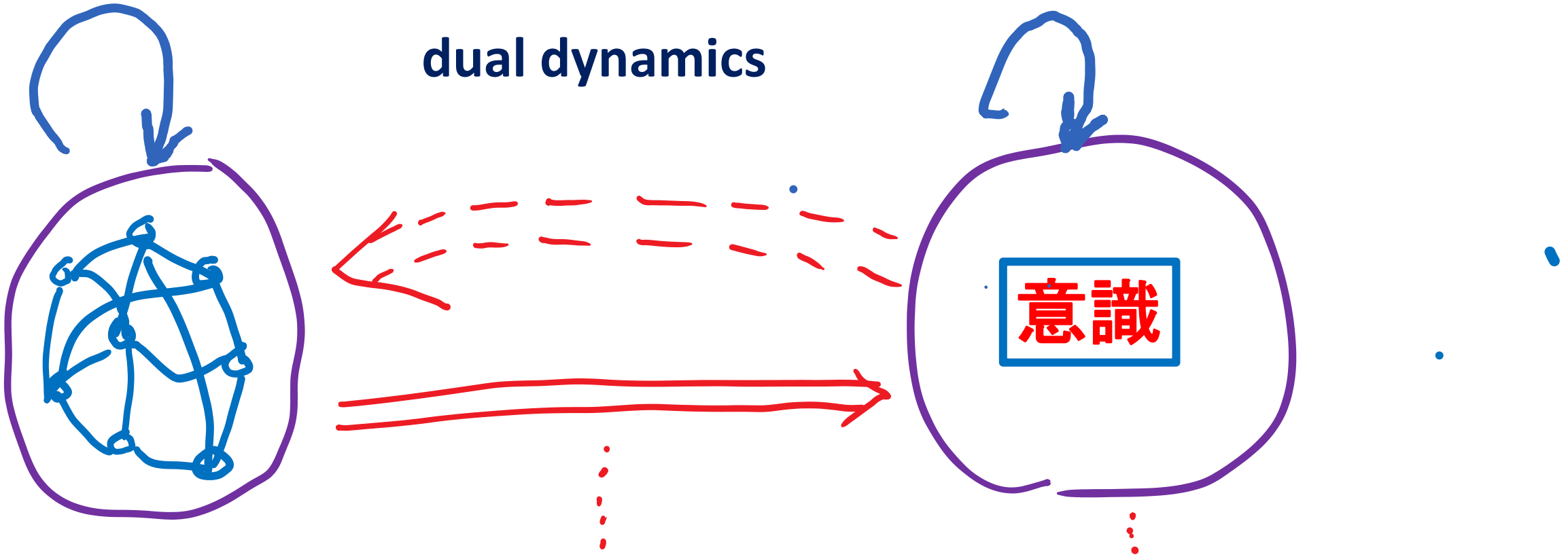
AI NN

意識の発生

共同作業、自分の意図を自分で知る

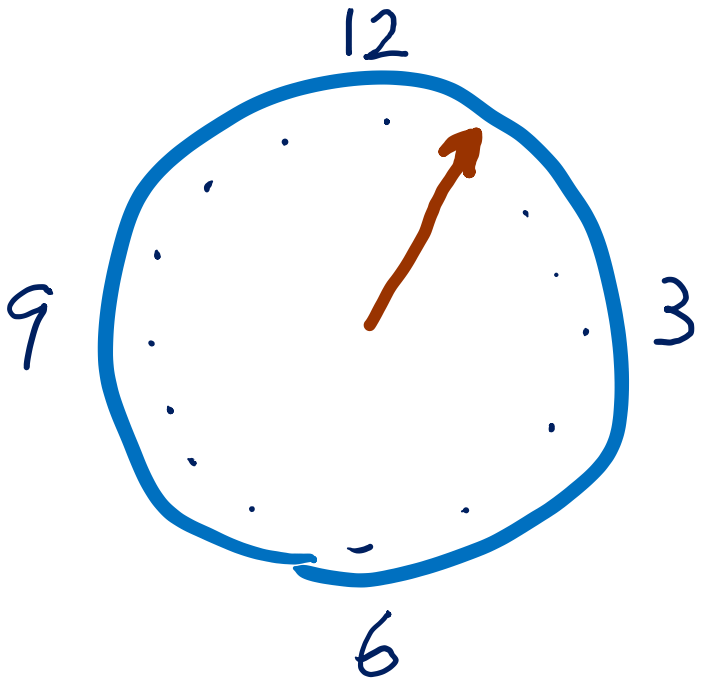
言語：　論理的思考、数学

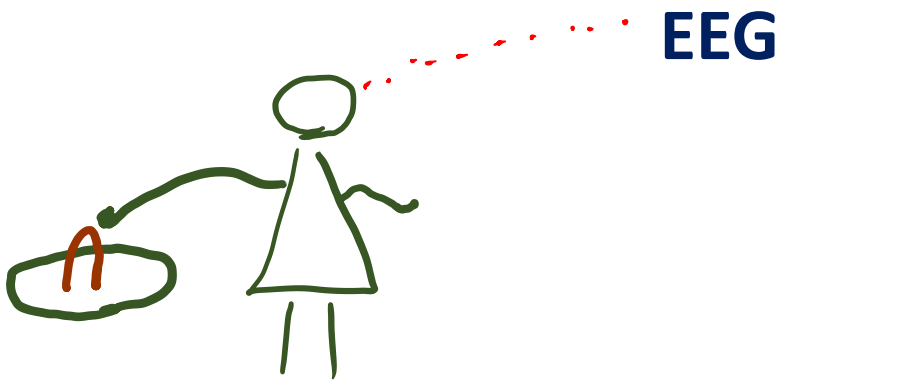
予測（先付け）と後付け　Prediction and Postdiction



ダイナミックス　　意思決定と行動　　反省、正当化、論理

Libet の実験：自由意志





When!

人工知能が脳に学ぶべきこと: 数理的理解

判断;制御;認知;記憶

意識と心の役割; 後付け

連想式記憶システム:知識体系

人工知能と倫理

人工知能の安全性、制御可能性

人工知能と戦争；人工知能の金融支配；
支配の道具、格差

暴走：　人間の暴走を範として

日本のAIの進むべき道：政府の戦略
ブームは終わる

超大国 ⟷ 文化国家

物量作戦はだめ
理論とアイデア
中小企業を含む現場との交流；産業の情報化

情報幾何の将来

Information Geometry Springer
諸科学の方法論
数学

結び: 若い研究者へ

自分のしたいことを忘れずに追及
当面の仕事だけではない一面従腹背

雑用の山、基礎研究、運をつかみ取れ