人工知能学会全国大会

脳、心、人工

付録ー統計神経力学の構想

理化学研究所　甘利俊一

人工知能の衝撃

人間の知的能力を超えるのか？
第４次産業革命

人工知能の歴史：
　記号と論理ー並列分散（ニューラルネット）

脳ー人間とは何か

思考・言語・意
人間・社会・文

宇宙誌と脳　－　脳ができるまで

ビッグバン	(138億年前)	物理学・化学
生命	(36億年前)	生命科学
脳・神経系	**(5億年前)**	神経科学・情報科学
文明・社会	(20万年前？)	脳科学・情報科学・人間科学

物質の法則：　　宇宙

生命の法則：　情報＋物質：　進化

文明の法則：　こころ＋情報＋物質：
社会・文化

人工知能と脳のモデル：一歴史の要約

第一次ブーム：　記号と論理　VS　パターンと学習

1956~　　AI　　　　　　　　脳モデル

Dartmous 会議　　　　　　　Perceptron

記号と論理　　　　　　　　学習する普遍計算機構

知的推論、ゲーム

実用的でない!!　　暗黒期（1965後半~1970’s）

第２次ブーム

1970~ AI
エキスパートシステム
(MYCIN, DENDRAL)

1980~ BT (神経回路)
MLP (backprop)
連想記憶モデル、ダイナミックス

一兆円産業か？

沈静化

確率Bayes推論
chess (1997)

第３次ブーム　2010~　　脳型の人工知能（融合）

深層学習　Deep learning

（畳み込み多層回路（福島）＋確率勾配降下：
日本でなぜ実現しな かったか？）

確率推論

深層学習の勝利 ーー人間以上の識別能力

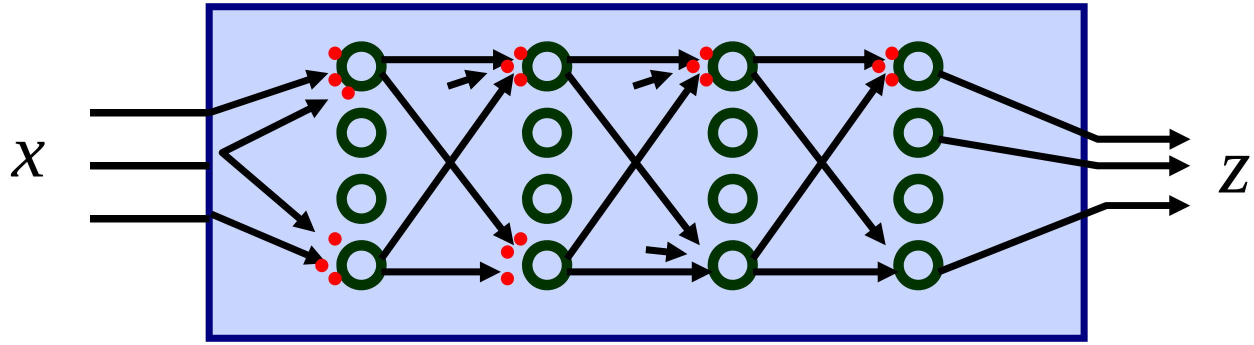
パターン認識: vision, auditory, sentence analysis

囲碁：強化学習

時系列とダイナミックス、動的パターン；言語処理

記号と論理　VS　パターンとダイナミックス、学習－融合

層状学習回路網
multilayer perceptron

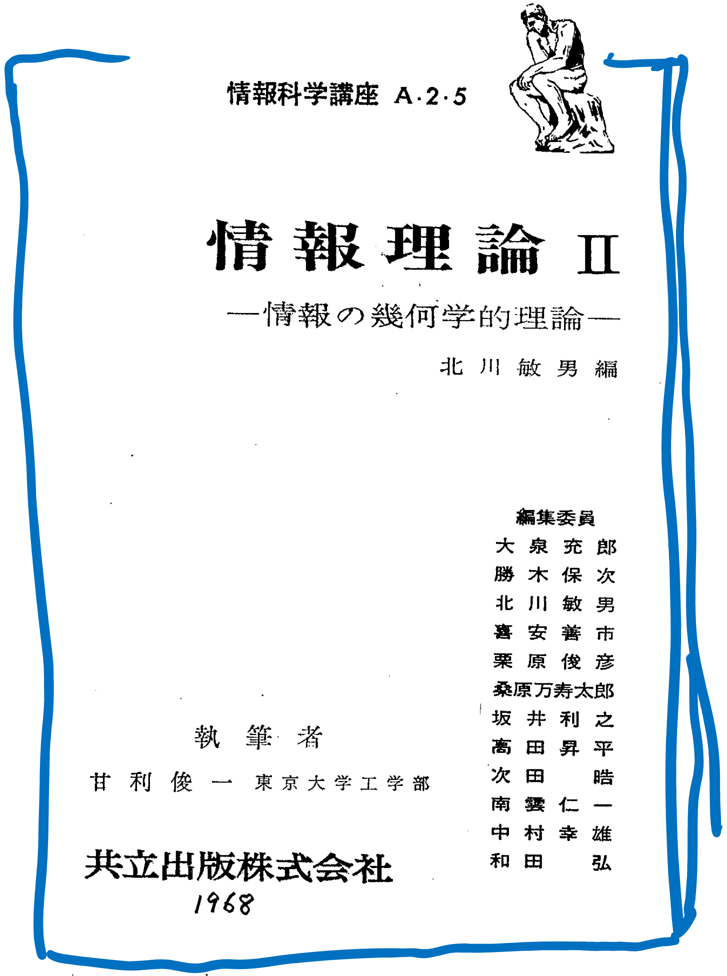


パーセプトロン　Perceptron
バックプロパゲーション　Backpropagation

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

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

最初のMLPの確率勾配降下学習法 (1967;1968)


情報科学講座 A·2·5

情報理論 II

—情報の幾何学的理論—

北川敏男 編

編集委員
大泉充郎
勝木保次
北川敏男
喜安善市
栗原俊彦
桑原万寿太郎
坂井利之
高田昇平
次田晧
南雲仁一
中村幸雄
和田弘

執筆者

甘利俊一 東京大学工学部

共立出版株式会社

1968

Information Theory II

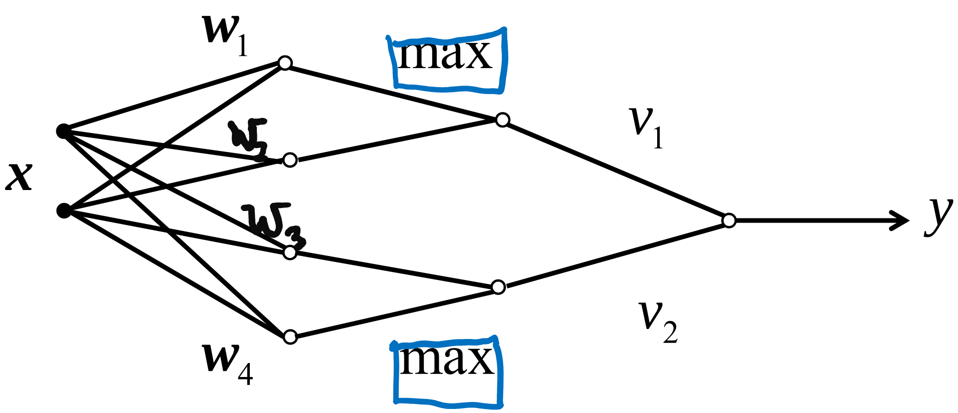
--Geometrical Theory of Information

Shun-ichi Amari
University of Tokyo

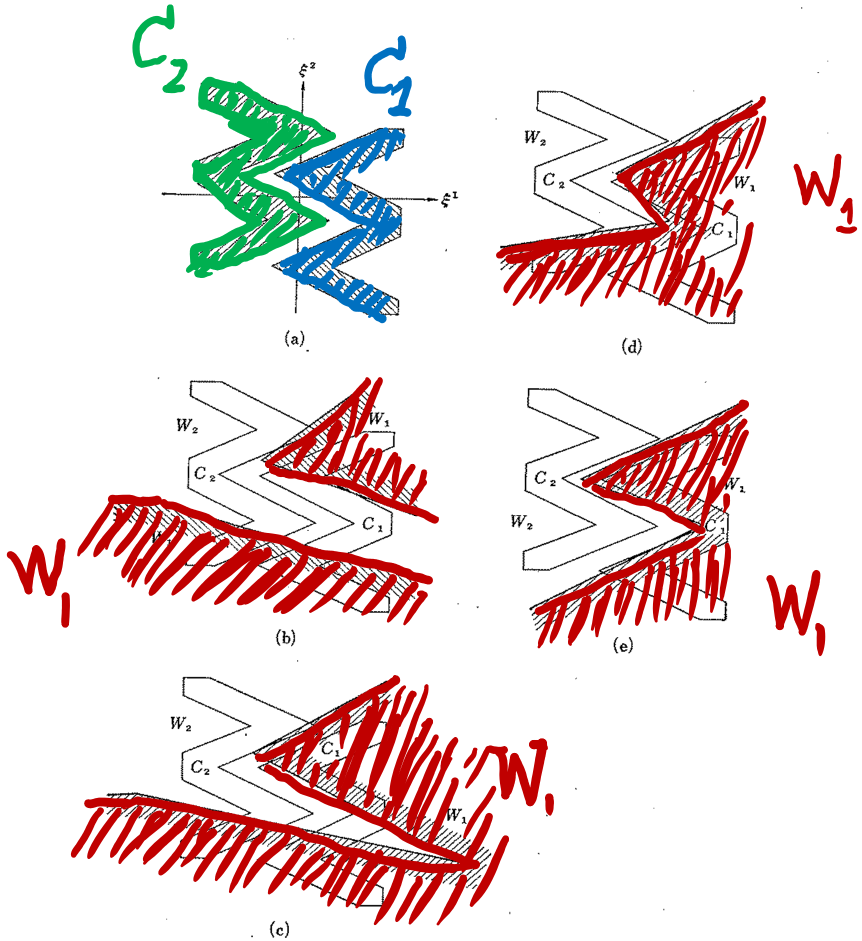
Kyouritu Press, Tokyo, 1968

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

アナログニューロン、シグモイド関数

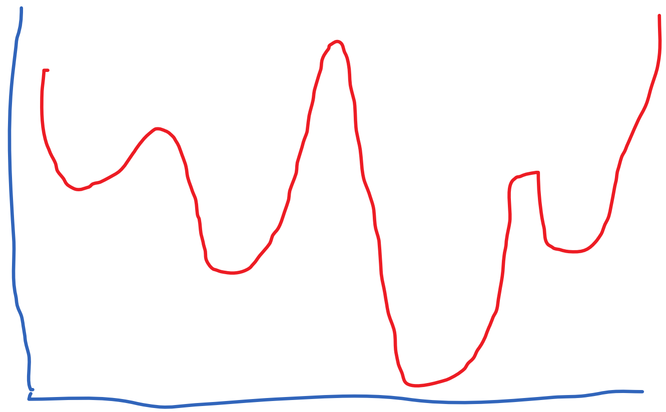


線形分離不可能パターン分類



深層学習

自己組織化学習（教師なし学習）
確率降下学習（教師あり学習）



情報表現の獲得

大域解と局所解：高次元

生成モデル　GAN　囲碁、言語：　何でもあり

大規模系の特徴

ランダム行列 A の固有値の分布

ほとんどが鞍点（極小解なし!!）

大規模回路

極小解は最小解の付近に集まる

深層学習

大量のデータ、計算力

入力を基に正解

現象の予測

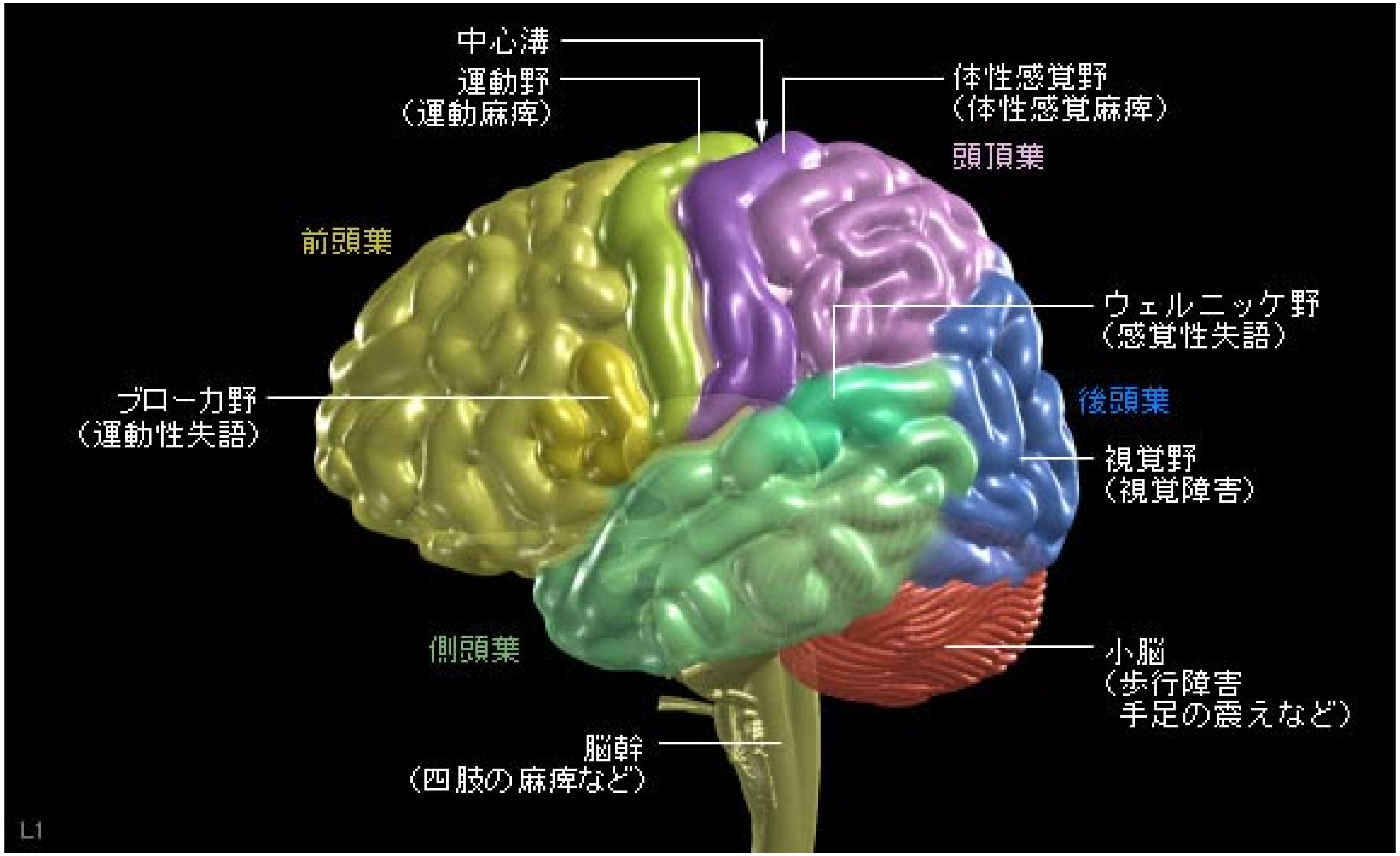
日蝕の予測

ケプラーの法則、ニュートン力学

原理の創出・理解ー人間

脳：大脳、海馬、小脳、脳幹

脳科学：ミクローマクロ、理論：神経回路網



数理脳科学は脳の基本原理を探求する

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）
→　神経科学（ありのままの脳）

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

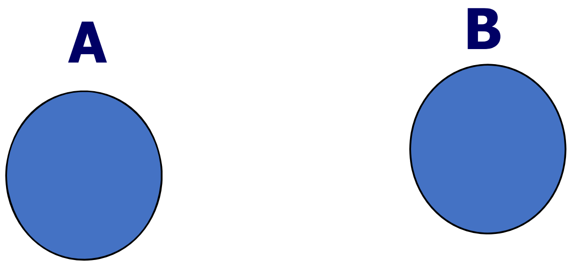
言語：　論理的思考、数学

心の理論

葛藤する心

究極のゲーム
（ultimatum game）

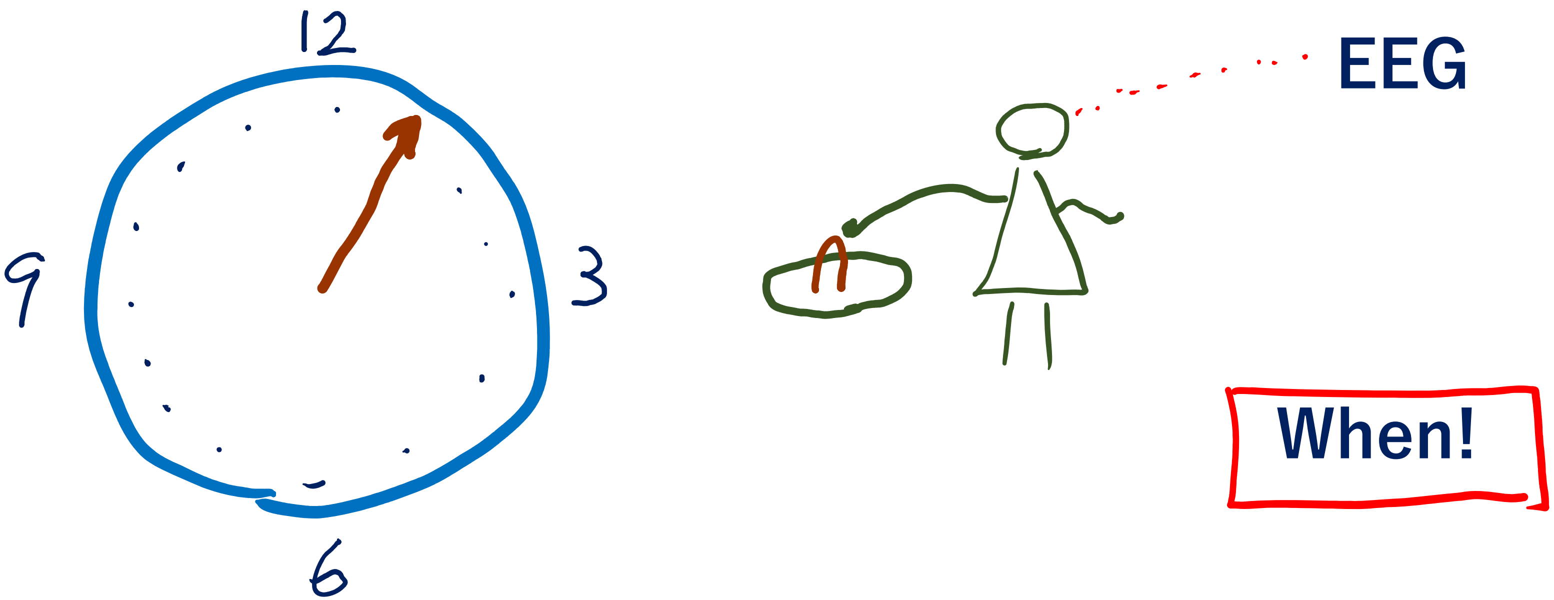
10万円



A:　配分を決定する----　7万円–3万円
B:　同意または拒否

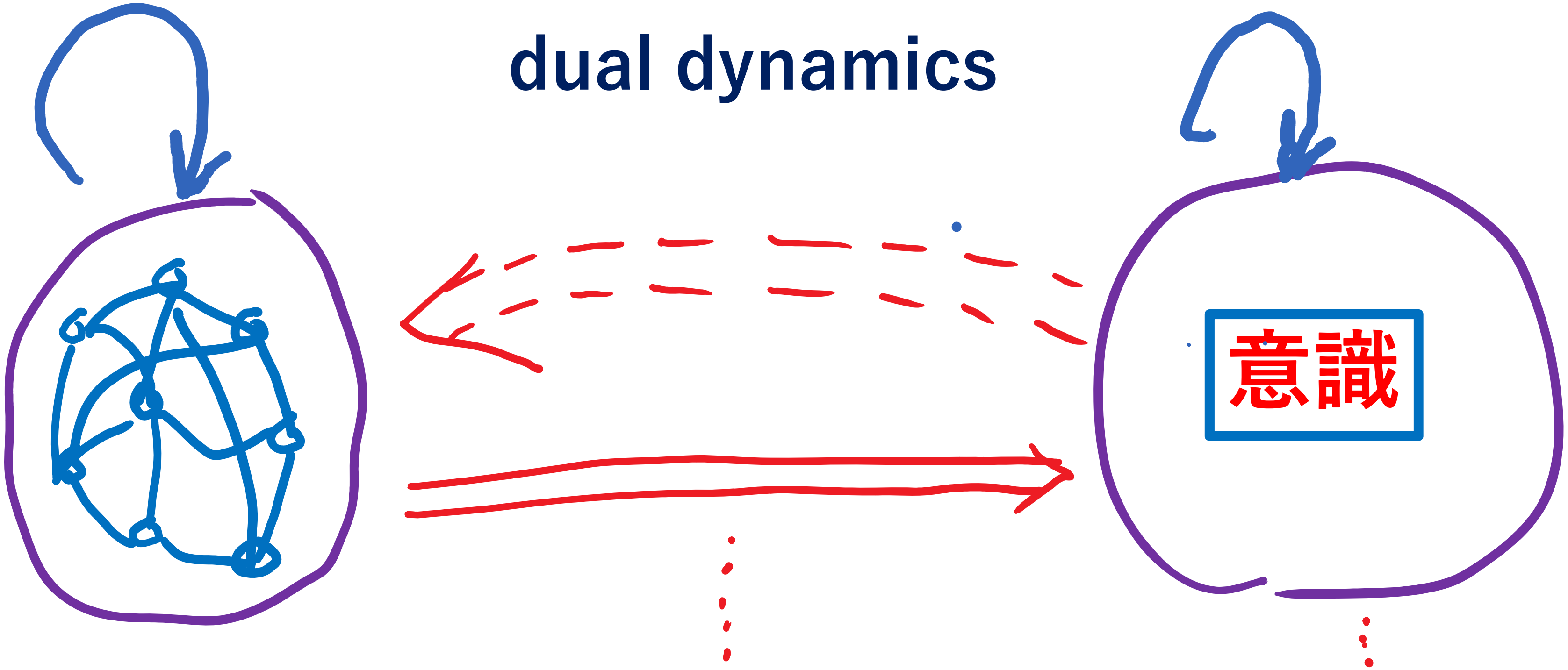
この時の脳の働き、各領野の確執
利益、公正、その後のこと、評判
二人か社会ゲームか

Libet の実験：自由意志



予測(先付け)と後付け prediction and postdiction

dual dynamics



ダイナミックス　　意思決定と行動　　反省、正当化、論理

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

心の理論

ロボットに意識はあるか

心を持ったロボットがつくれるか？

人の心の動きを理解する

ロボットが心を持つように見える
（感情移入）

ロボットが心を持てるのか？

人間の心：進化の産物

意識、意図、論理、感情

種の生存と個の不合理
人間は不合理；　芸術、喜び、愛、苦悩：使命感
ただ一度の、かけがいのない人生

ロボットは合理的

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

社会への影響

失業問題：人口減　：AIは仕事を奪うか？　より高度な仕事

格差の拡大：

ベーシックインカムと人類の家畜化：働く喜び

人工知能と技術的特異点　　　２０４５

人工知能が人間を超えるとき
　人工知能が研究し、技術を進める

　人間は素晴らしいが、愚かである。

人間はどんな知能システムを作るのか？
　　社会の進化と支配

人工知能と未来社会の設計

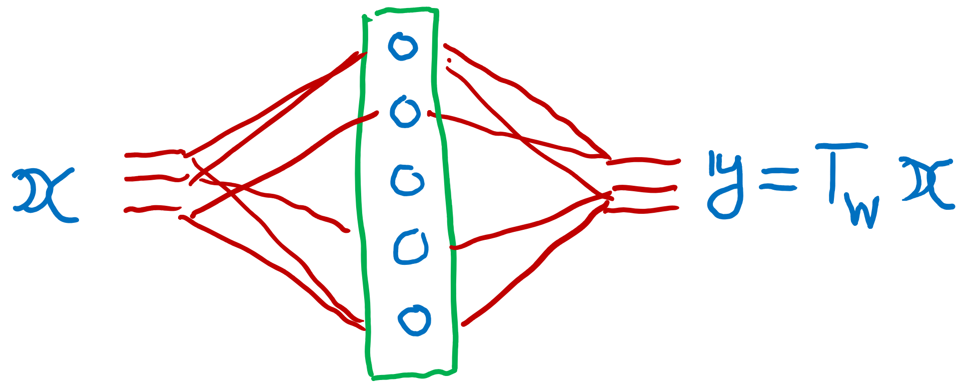
深層学習を超えて

科学研究、技術開発

社会、文明　　その脆弱性・崩壊

統計神経力学

Rozonoer (1969)
Amari (1971; 1974)
Amari et al (2013)
Toyoizumi et al (2015)
Poole, …, Ganguli (2016)
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

巨視変数

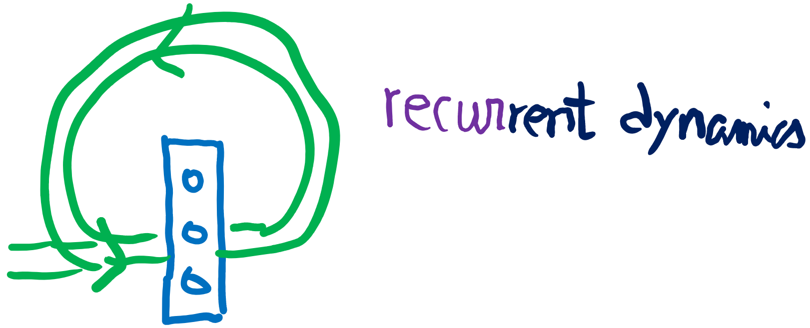
活動度 : $A = \dfrac{1}{n}\sum x_i^{\,2}$

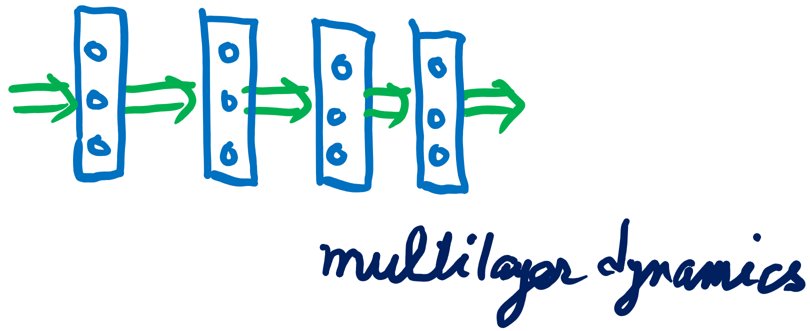
距離・計量: $D\!=\!D[\mathbf{x} : \mathbf{x}']$

曲率 :

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

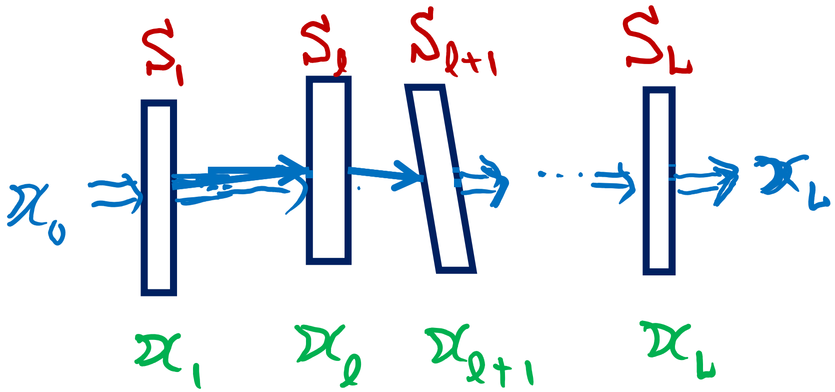
$$D_{l+1} = K(D_l)$$





深層回路

$$\underset{l+1}{x_i} = \varphi\left(\sum w_{ij} \underset{l}{x_i} + w_{0i}\right)$$



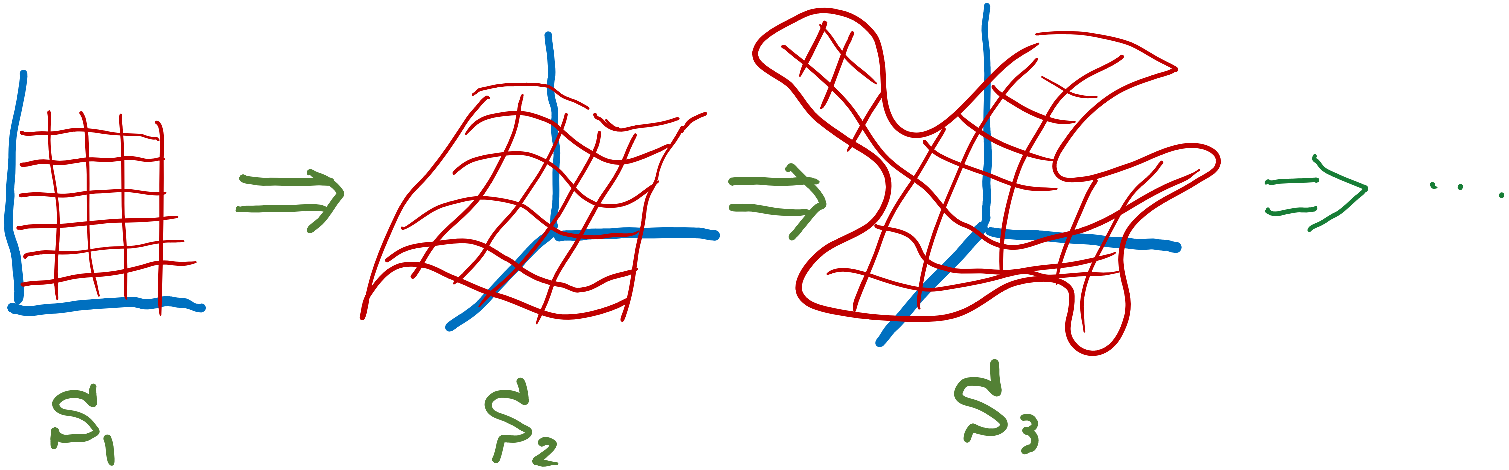
$$\underset{l}{A} = \frac{1}{n_l} \sum \underset{l}{x_i}^{2}$$

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

$$\underset{l+1}{A} = F(\underset{l}{A})$$

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

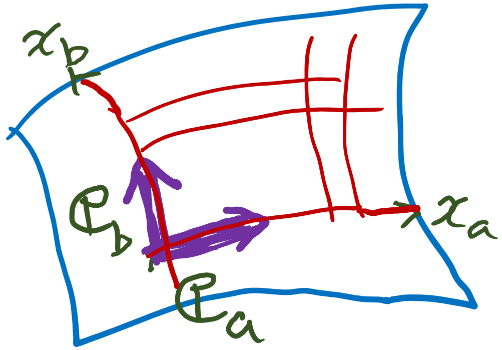
引き戻し計量（リーマン計量・距離・曲率）

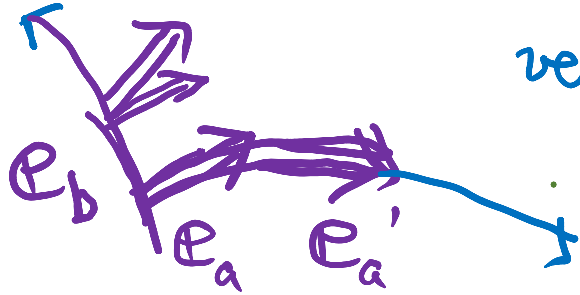


$$ds^2 = \sum g^{l}{}_{ab}\, dx^a dx^b = \frac{1}{n_l} d\mathbf{x}^l \cdot d\mathbf{x}^l$$

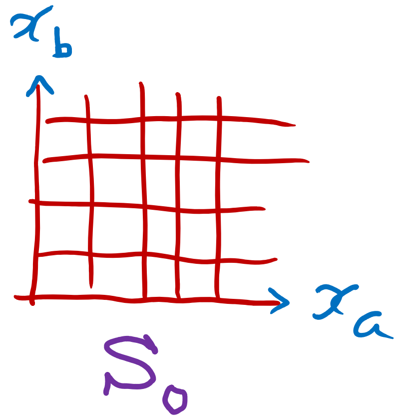
$$g^{l}{}_{ab} = \mathbf{e}^{l}{}_{a} \cdot \mathbf{e}^{l}{}_{b}$$

曲率



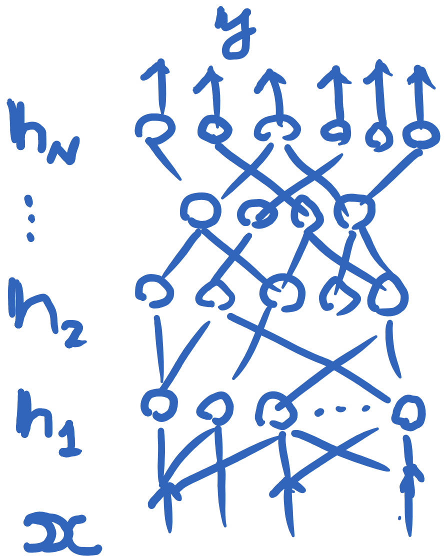


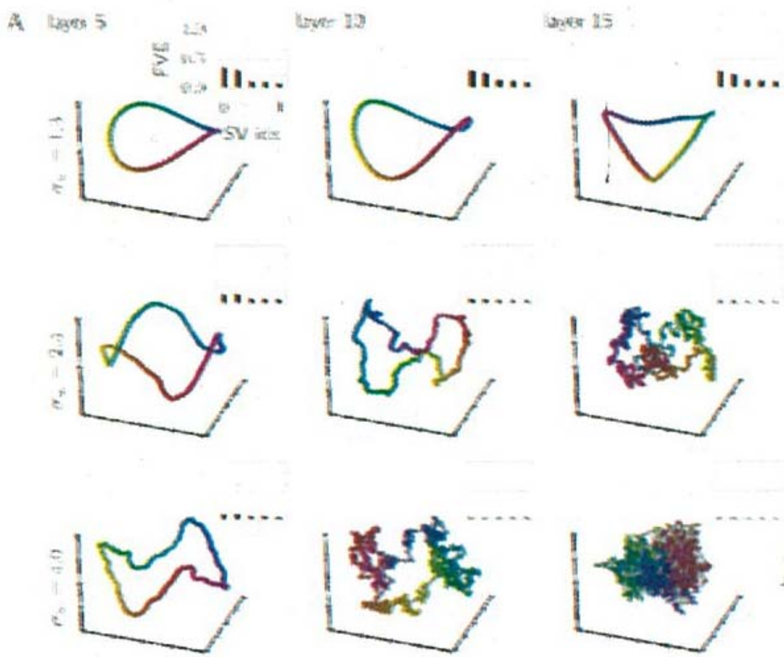
rotation
vertical



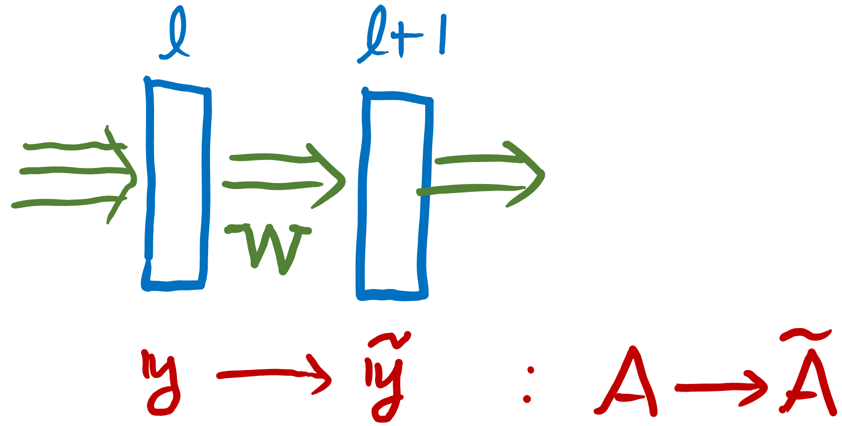
$$\overset{\ell}{H}_{abi} = \nabla_a \overset{\ell}{\mathbf{e}}_b$$

Poole et al (2016)
Random deep neural networks





活動度の力学



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

$$u_\alpha \sim N(0, \sigma_A^2); \qquad \sigma_A^2 = \sigma^2 A + \sigma_b^2$$

$$\tilde{A} = \frac{1}{n_{l+1}} \sum (\tilde{y}_\alpha)^2 = E[\varphi(u_\alpha)^2] = \chi_0(A) = \frac{1}{2\pi} \cos^{-1}(-\frac{\sigma_A^2}{1+\sigma_A^2})$$

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

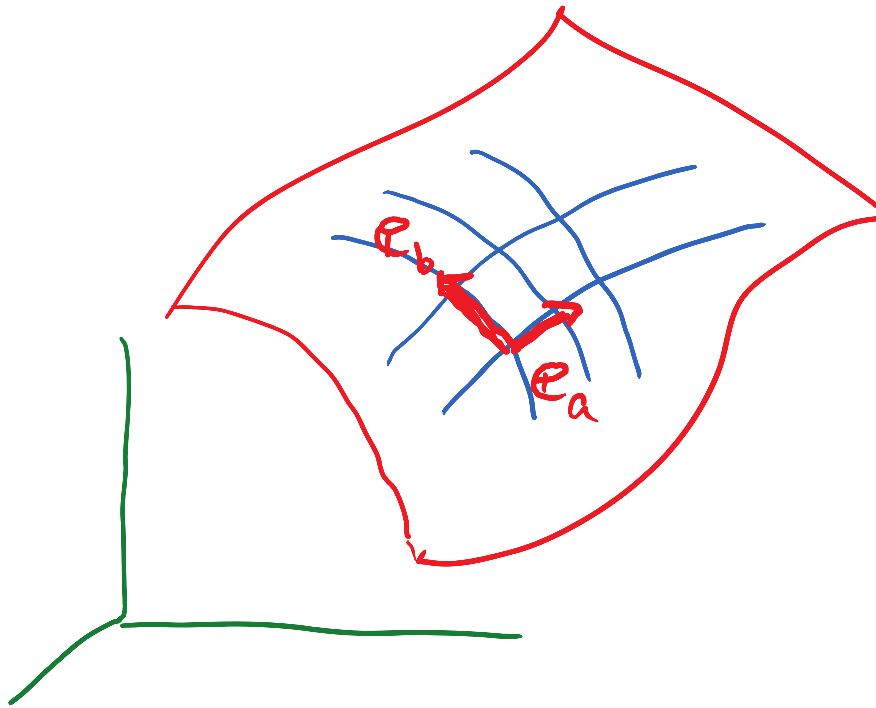
Basis vectors

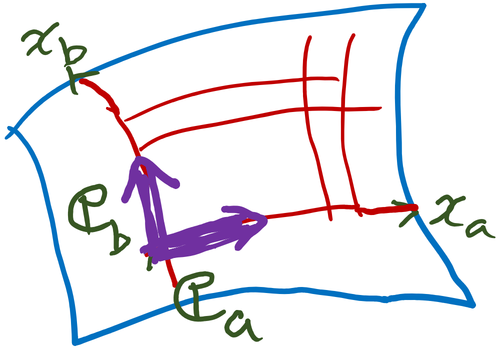
$$dx_{i_l} = \sum \varphi'\left(u_{i_l}\right) W^{i_l}_{i_{l-1}} dx_{i_{l-1}} = \sum B^{i_l}_{i_{l-1}} dx_{i_{l-1}}$$

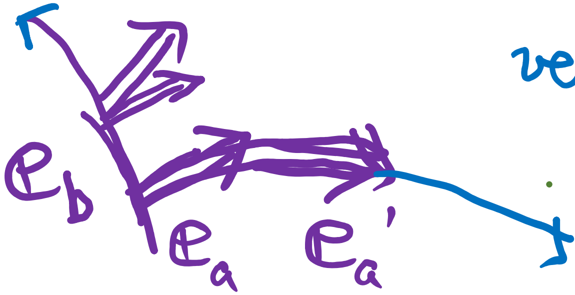
$$d\overset{l}{\boldsymbol{x}} = \overset{l}{B}\, d\overset{l-1}{\boldsymbol{x}} \qquad \left(= \overset{l}{B}\dots\overset{m}{B}\, d\overset{m-1}{\boldsymbol{x}}\right)$$

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

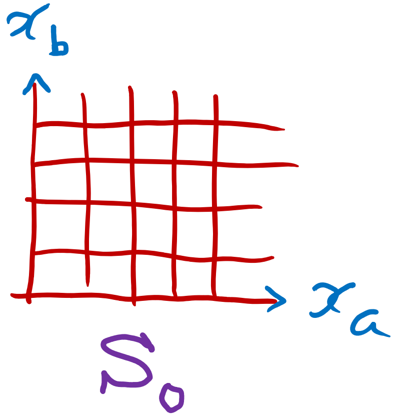
$$\overset{l}{\boldsymbol{e}}_a = \overset{l}{B}\,\overset{l-1}{\boldsymbol{e}}_a = \overset{l}{B}\dots\overset{m}{B}\,\overset{m-1}{\boldsymbol{e}}_a$$







rotation
vertical



$$\overset{\ell}{g}_{ab} = \frac{1}{n_l}\overset{\ell}{\mathbf{e}}_a \cdot \overset{\ell}{\mathbf{e}}_b$$

Mectric

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

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

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

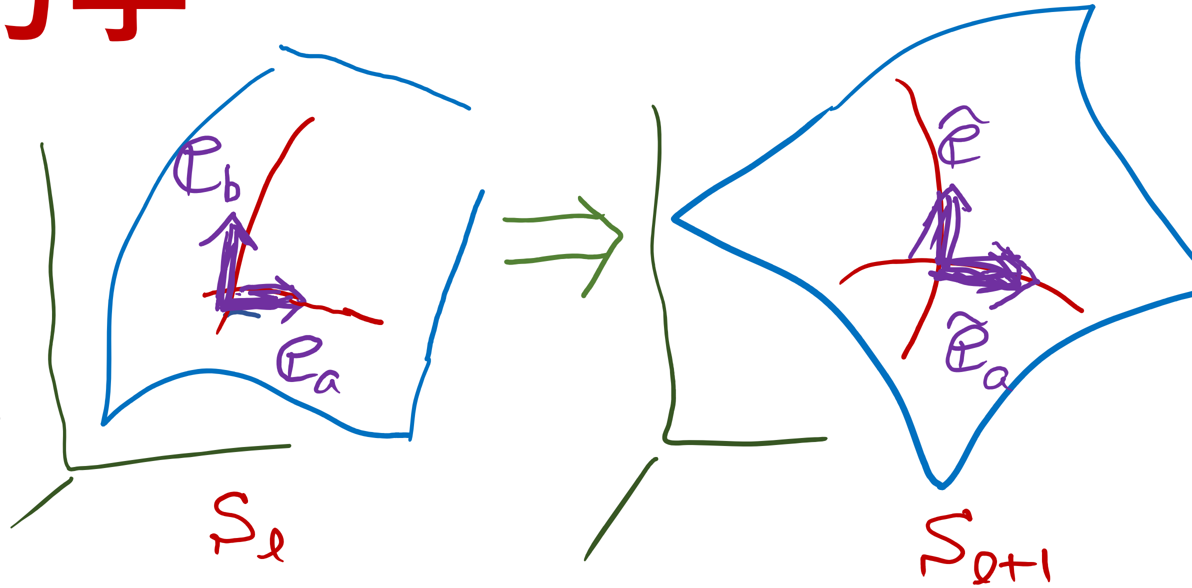
$$\chi_1 = \sigma_l^2 E\left[\varphi'\left(u_{i_l}\right)^2\right]$$

リーマン計量の力学

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

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

$$ds^2 = \sum g_{ij} dy^i dy^j = < d\mathbf{y}, d\mathbf{y} >$$



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

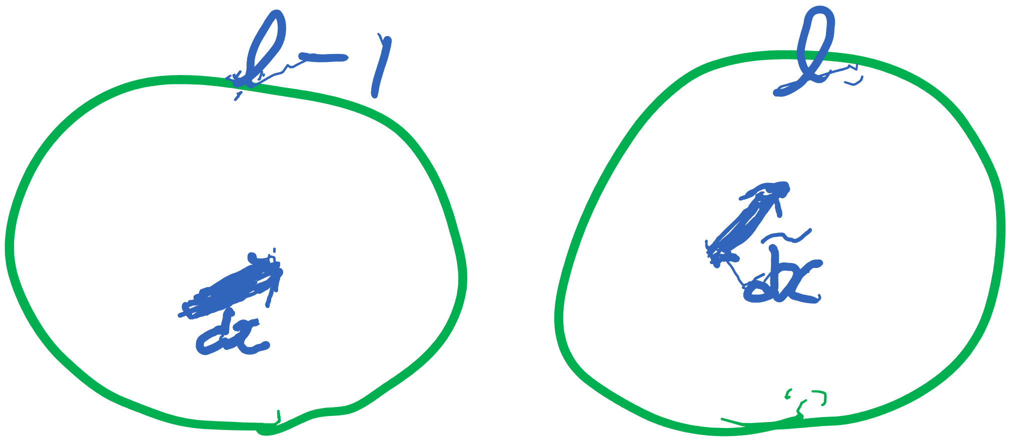
$$< \tilde{\mathbf{e}}_a, \tilde{\mathbf{e}}_b > = \tilde{g}_{ab} = \sum B_k^\alpha B_j^\alpha < \mathbf{e}_k, \mathbf{e}_j >$$

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

平均場近似

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

law of large numbers



$$d\,s^2 = \sum (dx^i)^2$$

$$d\,\overset{l}{s}^2 = \sum BBd\overset{l-1}{x}d\overset{l-1}{x}$$

$$\sum B^{i_l}_{i_{l-1}} B^{i_l}_{i'_{l-1}} = \sum [\varphi'(u_{i_l})^2 W^{i_l}_{i_{l-1}} W^{i_l}_{i'_{l-1}}]$$

$$= n\mathrm{E}[\varphi'(u_{i_l})^2 W^{i_l}_{i_{l-1}} W^{i_l}_{i'_{l-1}}] + \mathrm{O}_p(1/n)$$

$$= \mathrm{E}[\varphi'(u_{i_l})^2]\mathrm{E}[W^{i_l}_{i_{l-1}} W^{i_l}_{i'_{l-1}}] = \chi\delta_{i_{l-1}i'_{l-1}} + \mathrm{O}_p(1/n)$$

$$\chi = \sigma^2 \int \varphi'(u_{i_l})^2 Dv$$

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

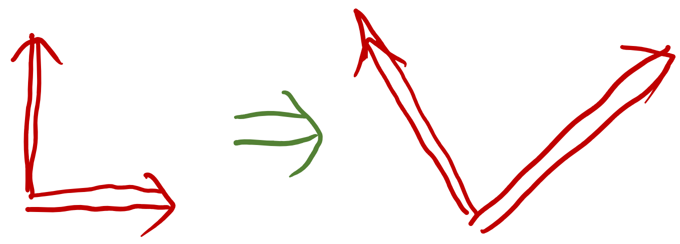
conformal transformation!

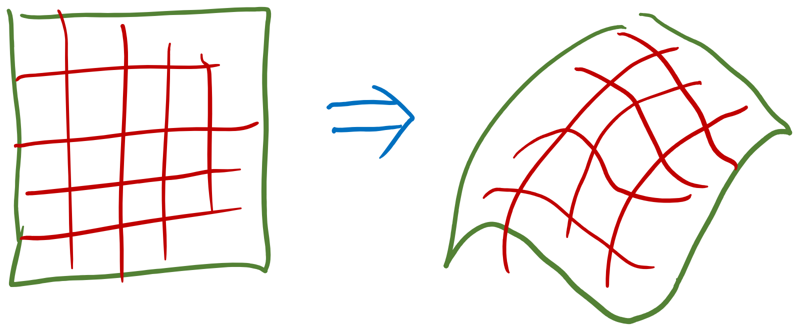
$$\bar{\chi}_1 = \bar{\chi}_1(\bar{A}) > 1:$$

拡大（カオス、Lyapnov指数）

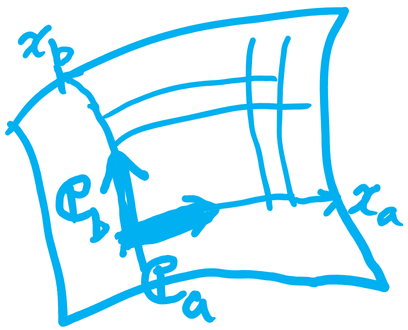
$$\Rightarrow g^{l}{}_{ab} = \Pi \chi_1 (A^{s}) \delta_{ab}$$

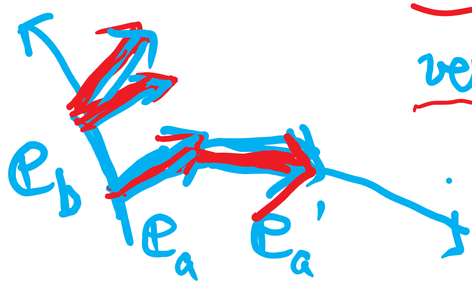
rotation, expansion





曲率の力学





rotation

vertical

$$\tilde{H}^{\alpha}_{ab} = \nabla_a \tilde{\mathbf{e}}^{l}_{b} = \partial_a \partial_b \tilde{y}^{\alpha}$$

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

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

Euler-Schouten曲率
Affine connection

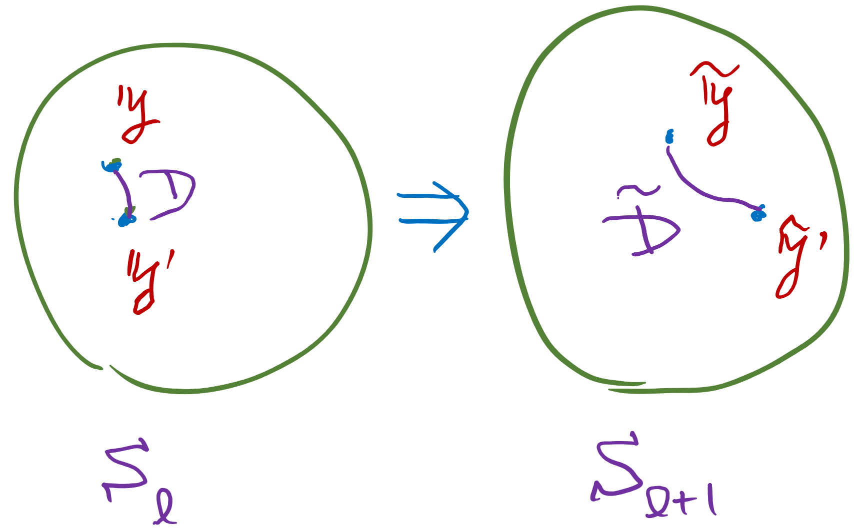
$$\tilde{H}^{2}_{ab} = |\tilde{\mathbf{H}}_{ab}|^{2}$$

距離法則 (Amari, 1974)

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

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

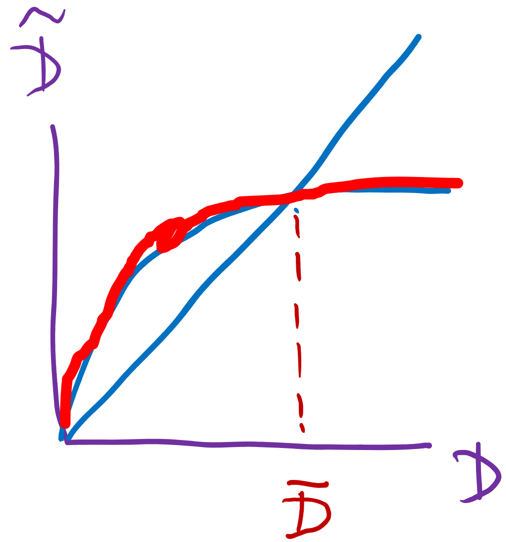
$$D = A + A' - 2C$$



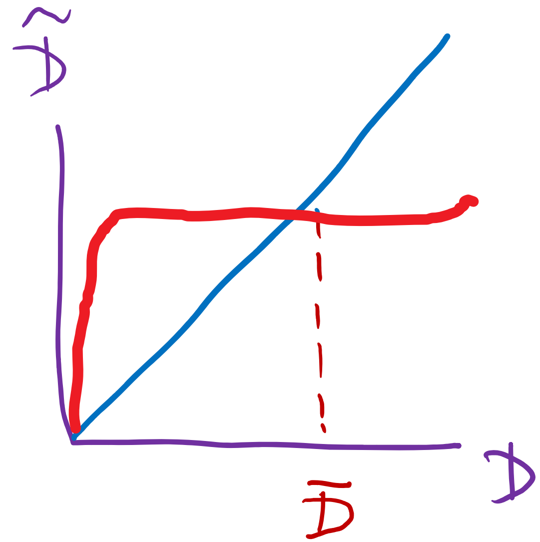
$$D_{l+1} = \xi(D_l)$$

$$\frac{d\tilde{D}}{dD} = \chi_1 > 1 \qquad D(\mathbf{x}, \mathbf{x}') \Rightarrow \overline{D}$$

$$\overline{D} = \xi(\overline{D})$$



$$D_l = \xi * \xi * \ldots \xi(D_0)$$



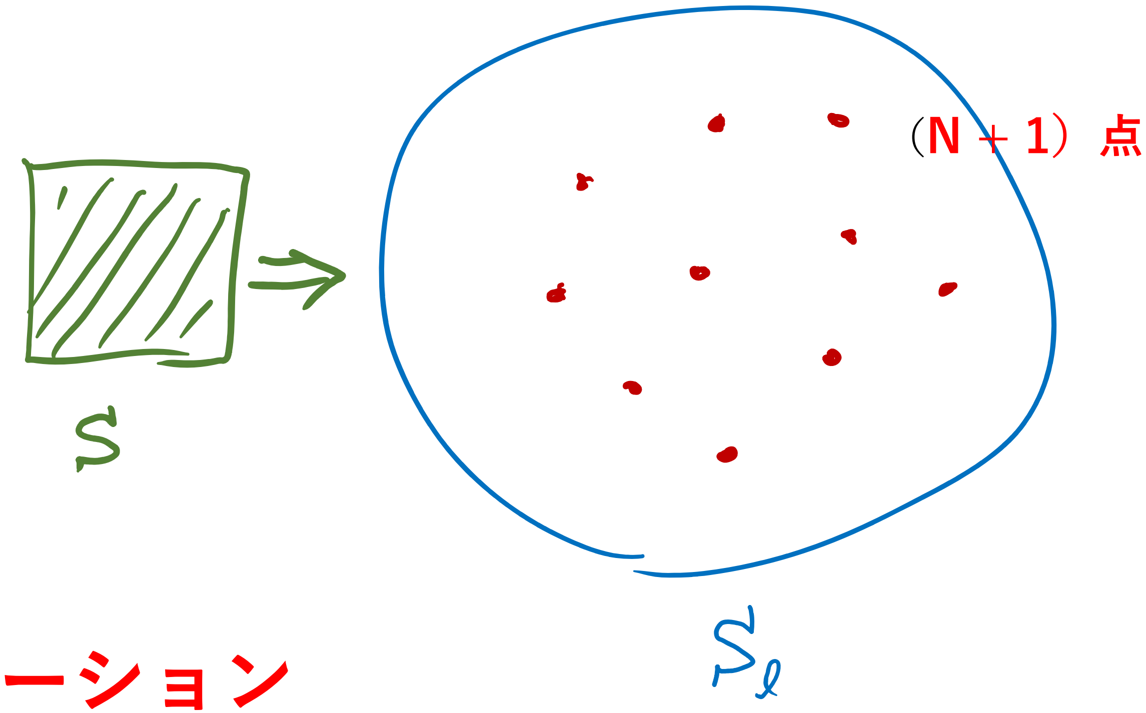
本当か! $n \to \infty; l \to \infty$

equidistance

$$D(\mathbf{x}_l, \mathbf{x}'_l) \to \bar{D}$$

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

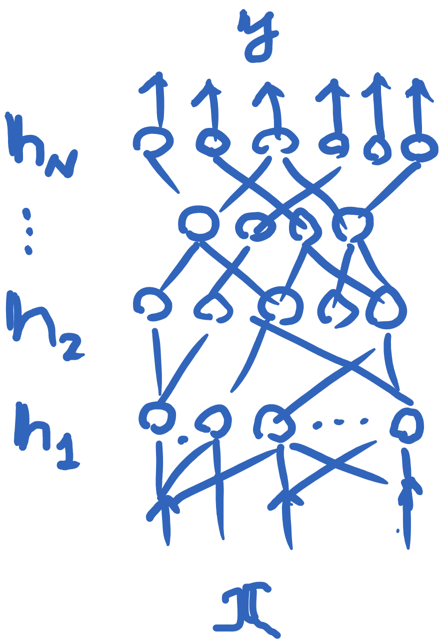
等距離

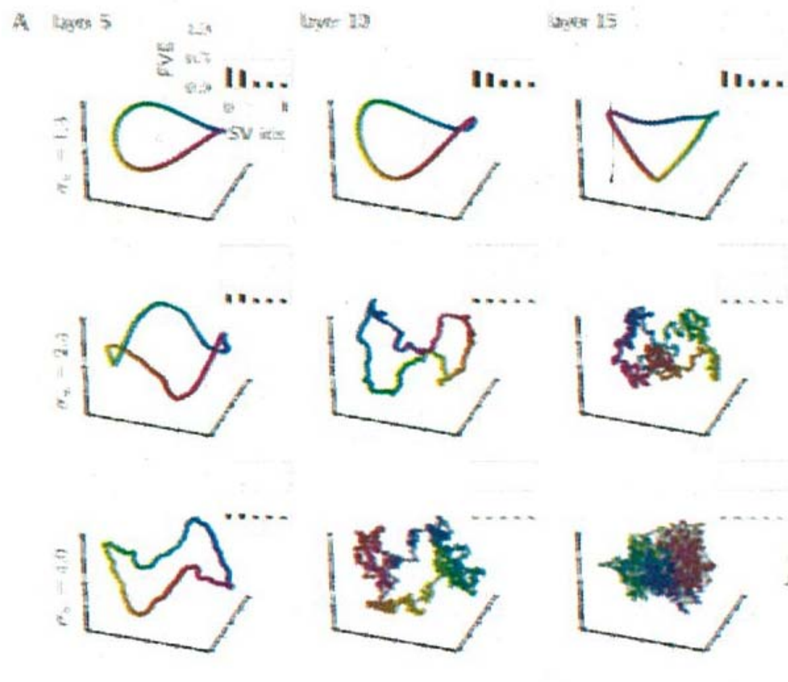


フラストレーション
フラクタル

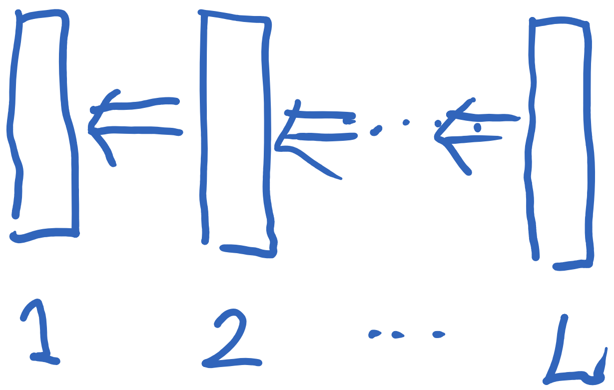
Poole et al (2016)
Deep neural networks

フラクタル
敵対的例題





Backward path; error back-propagation
Fisher information matrix



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

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

Stochastic model：　深層回路の多様体

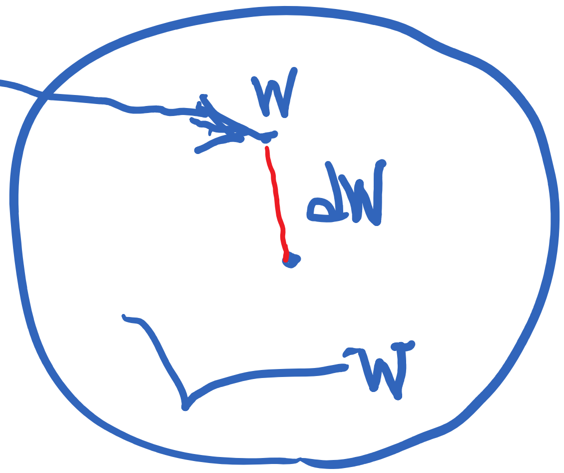
$$y=\varphi(u)+\varepsilon; \qquad \varepsilon \sim N(0,1)$$

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

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

$$ds^2=dWGdW$$

Fisher information



Riemannian

Fisher information

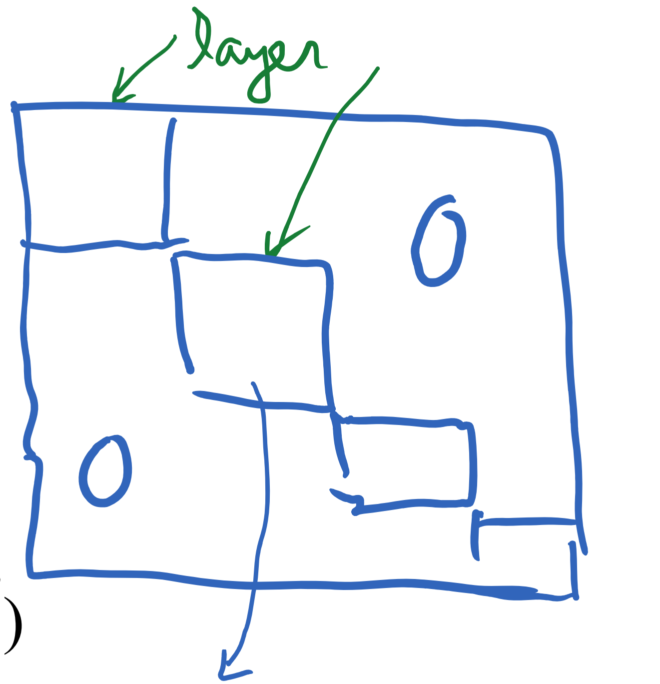
$$G = E_x\left[\frac{\partial \varphi}{\partial W_m}\frac{\partial \varphi}{\partial W_l}\right]$$

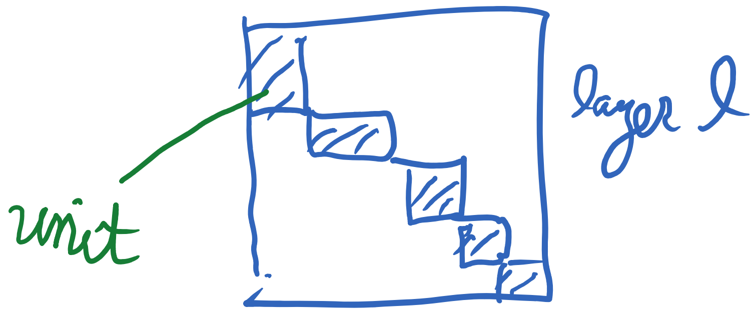
$$\frac{\partial \varphi^{l}}{\partial W_m} = \overbrace{\varphi' W}^{B}\frac{\partial \varphi^{l-1}}{\partial W_m} = B\frac{\partial \varphi^{l-1}}{\partial W_m} = \overbrace{BB\ldots B}^{B}\frac{\partial \varphi^{m+1}}{\partial W_m}$$

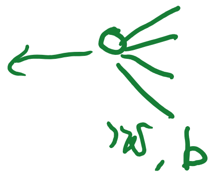
$$G\left(W_l, W_m\right) = \prod \chi_1 E_x\left[\varphi'\left(\overset{l}{\mathbf{w}}_i\right)^2 \overset{l-1}{\mathbf{x}}\ \overset{l-1}{\mathbf{x}}\right] + O_p(1/\sqrt{n})$$

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

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







$$\mathbf{B}_{i_m}^{i_l} = \sum B_{i_{l-1}}^{i_l} B_{i_2}^{i_{l-1}} \ldots\ldots B_{i_{m-1}}^{i_m} = O_p\left(\frac{1}{n}\right)$$

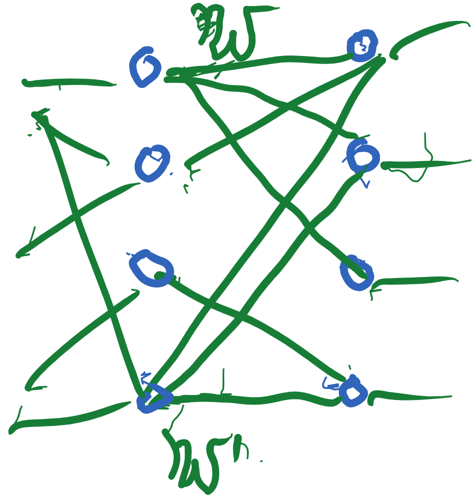
$$\sum \mathbf{B}_{i_m}^{i_l} \mathbf{B}_{i'_m}^{i_l} = \chi\chi \cdot\cdot \chi\delta_{i_m i'_m} + O_p\left(\frac{1}{n}\right)$$

$$G(W_l, W_m) = \mathrm{E}_x[\sum \mathbf{B}^{i_L}_{i_l} \mathbf{B}^{i_L}_{i_m} \varphi' \varphi' \mathbf{xx}]$$

$$+\mathrm{o}_p(1/n); \quad l \neq m$$

$$G(W_l, W_l) = \mathrm{E}_x[\sum \mathbf{B}^{i_L}_{i_l} \mathbf{B}^{i_L}_{i'_m} \varphi' \varphi' \mathbf{xx}]$$

$$= \delta_{i_l i'_l} \mathrm{E}_x[\varphi' \varphi' \mathbf{xx}]_{i_{l-1} i'_{l-1}}$$



$G(w, w')$

Domino Theorem

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

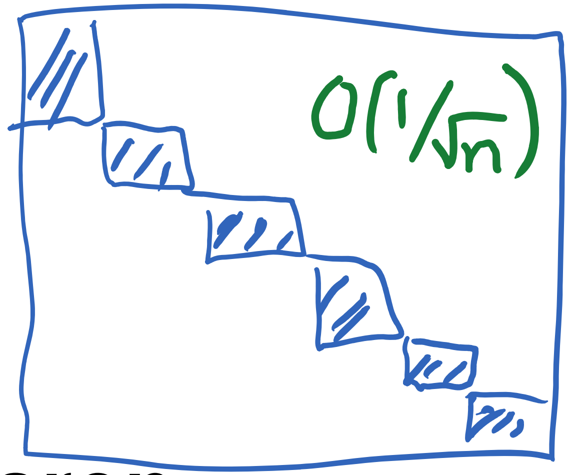
$$\frac{\partial \boldsymbol{x}^{l}}{\partial W^{m}} = B\frac{\partial \boldsymbol{x}^{l-1}}{\partial W^{m}} = BB\frac{\partial \boldsymbol{x}^{l-2}}{\partial W^{m}} = BBB\cdots$$

$$\Sigma\delta_{i_L i'_L} B^{i_L}_{i_{L-1}} B^{i'_L}_{i'_{L-1}} = \chi_1 \delta_{i_{L-1} i'_{L-1}} + O_p\left(\frac{1}{\sqrt{n}}\right)$$

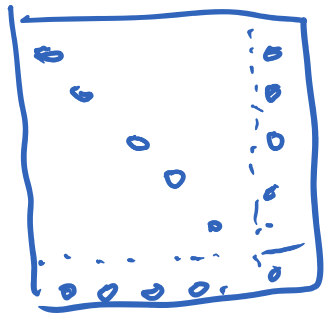
$$\Sigma\delta_{i_L i'_L} B^{i_L}_{i_{L-1}} B^{i'_L}_{i'_{L-1}} BBBB = \chi_1\chi_1\chi_1\chi_1 \delta_{i_{L-1} i'_{L-1}} + O_p\left(\frac{1}{\sqrt{n}}\right)$$

Unitwise natural gradient

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

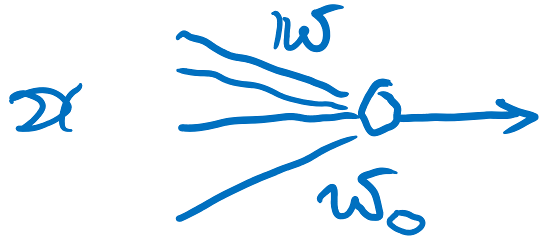
$G^{-1} =$ 

Y. Ollivier; Marceau-Caron



Fisher information of unit

$\tilde{w} = (w, w_0)$



$$y = \varphi\left(\boldsymbol{w} \cdot \boldsymbol{x} + w_0\right)$$

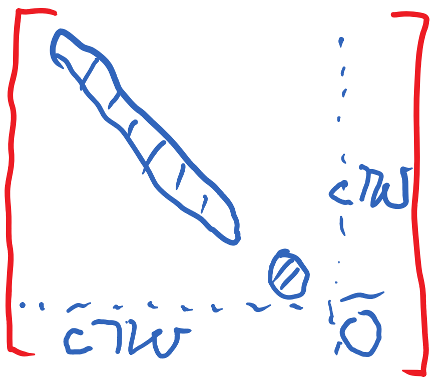
$$G = \mathrm{E}_{\boldsymbol{x}}\left[\partial_{\boldsymbol{w}}\varphi\ \partial_{\boldsymbol{w}}\varphi\right] = \mathrm{E}_x[(\varphi')^2 \mathbf{xx}]$$

$G(\tilde{w}, \tilde{w})$

$$\{\mathbf{e}_1, \mathbf{e}_2, \ldots \mathbf{e}_n\} \Rightarrow \{\mathbf{e}_1{}^*, \mathbf{e}_2{}^*, \ldots, \mathbf{e}_n{}^*\} : \text{ortho-normal basis}$$

$$\mathbf{e}_n{}^* = \frac{\mathbf{w}}{w},$$

Input x: independent and identically distributed, 0-mean



$$G = A\mathbf{I} + \frac{B}{\boldsymbol{w}^2}\boldsymbol{w}\boldsymbol{w} + \frac{C}{\boldsymbol{w}}\left(\boldsymbol{w}\boldsymbol{b} + \boldsymbol{b}\boldsymbol{w}\right)$$

$$\boldsymbol{b} = \begin{bmatrix} 0 & 0 & \cdots & 0 & 1 \end{bmatrix}'$$

$$\underbrace{A, B, C}_{(w,\, b)}$$

G^{-1} : similar form

$$\Delta \mathbf{w} = -\eta e G^{-1} \mathbf{x}$$

TANGO

QD自然勾配法?

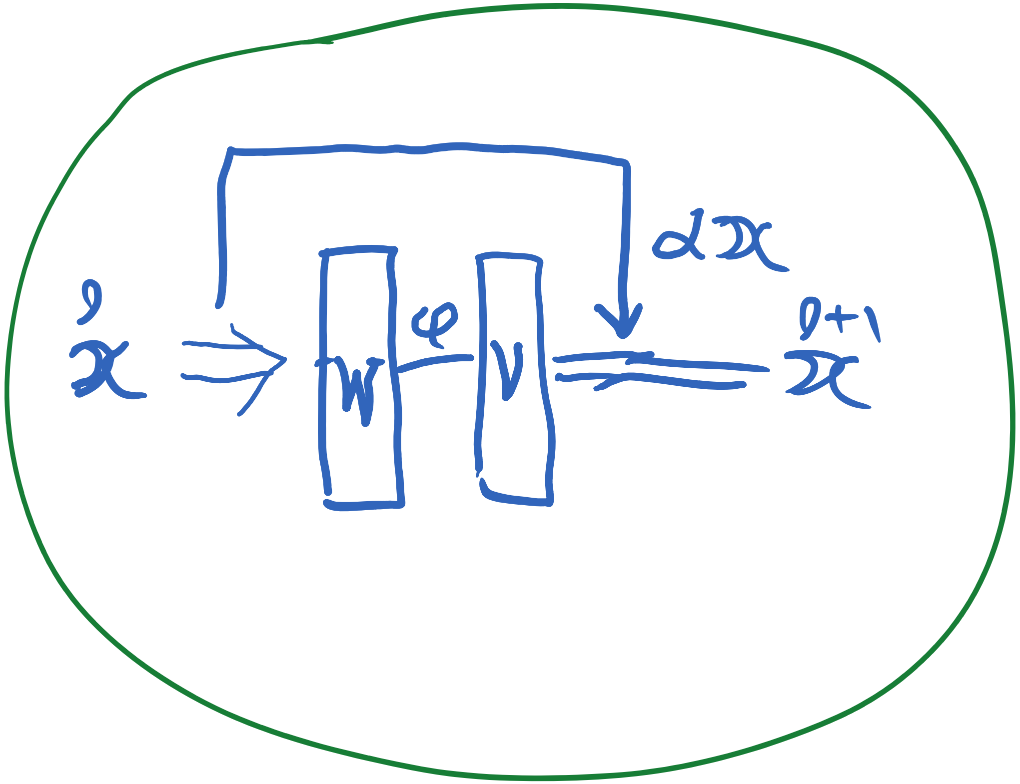
$$= -\eta e[\overline{A}\mathbf{x} + (\frac{\overline{B}}{w^2}\mathbf{w} \cdot \mathbf{x} + c)\mathbf{w}]$$

$$\Delta w_0 = -\eta e D$$

Lesnet

$$\overset{l}{\boldsymbol{x}} = V\varphi\left(W\overset{l-1}{\boldsymbol{x}}\right) + \alpha\overset{l-1}{\boldsymbol{x}}$$

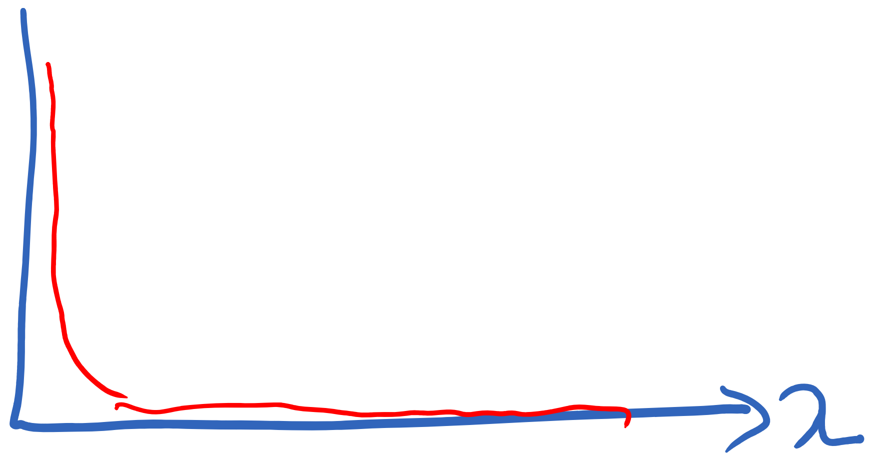
$$\chi_1 \to \sigma_v^2 \chi_1 + \alpha^2$$



Karakida theory

eigenvalues of G

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



distorted Riemannian metric