知識転移の学習曲線:
可解な学習ダイナミクスの知見

唐木田 亮
産業技術総合研究所 人工知能研究センター

Workshop OT 2023 「最適輸送とその周辺」
2023-03-15

自己紹介

2008-2012 東京大学 理学部物理学科
2012-2017 修士/博士課程
東京大学 新領域創成科学研究科
複雑理工学専攻 (指導教員: 岡田真人 教授)
興味: 非線形動力学, 脳の神経回路

2017- 研究員 産業技術総合研究所 人工知能研究センター
機械学習研究チーム @お台場
2021- 主任研究員

ニューラルネット・機械学習の典型的/普遍的な挙動. 力学的/統計力学的視点.

"**Information Geometry Connecting Wasserstein Distance and Kullback-Leibler Divergence via the Entropy-Relaxed Transportation Problem**", Amari, K, Oizumi, Information Geometry, 2018

"Fisher Information and Natural Gradient Learning in Random Deep Networks", Amari, K, Oizumi, AISTATS, 2019

Wasserstein距離と情報幾何

[Amari, K & Oizumi, Info. Geo., 2018]

周辺分布: $\boldsymbol{p}, \boldsymbol{q} \in S_{n-1}$

エントロピー正則化付き最適輸送コスト [Cuturi, '13] :

$$\varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) = \min_{P \in \Pi(\boldsymbol{p}, \boldsymbol{q})} \frac{1}{1+\lambda} \sum_{i,j} M_{ij} P_{ij} + \frac{\lambda}{1+\lambda} \sum_{ij} P_{ij} \ln P_{ij}$$

$\boldsymbol{\eta} = (\boldsymbol{p}, \boldsymbol{q})$に関して凸関数　→ 双対平坦構造を定める

- Sinkhornアルゴリズムはe-射影の繰り返し
- Divergenceを作る
  - λ-divergence (canonical divergence)

$$D_\lambda[\boldsymbol{p} : \boldsymbol{q}] = \varphi_\lambda(\boldsymbol{p}, \boldsymbol{p}) - \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) - \nabla_{\boldsymbol{q}} \varphi_\lambda(\boldsymbol{p}, \boldsymbol{q}) \cdot (\boldsymbol{p} - \boldsymbol{q}) = \frac{\lambda}{1+\lambda} KL\left[P_\lambda(\boldsymbol{p}, \boldsymbol{p}) : P_\lambda(\boldsymbol{p}, \boldsymbol{q})\right]$$

$$\lambda \to \infty \nearrow \quad KL[\boldsymbol{p} : \boldsymbol{q}]$$

  - λ→0でWasserstein距離とつながる例 [Amari, K, Oizumi & Cuturi, N.C., 2019]

Outline

- 研究背景
  ・ニューラルネットの統計力学的解析
  ・アルゴリズムと解選択
  NTK, 対角線形ネット

- 継続学習における知識転移
  - 汎化誤差のレプリカ解析
  - 自己知識転移と忘却

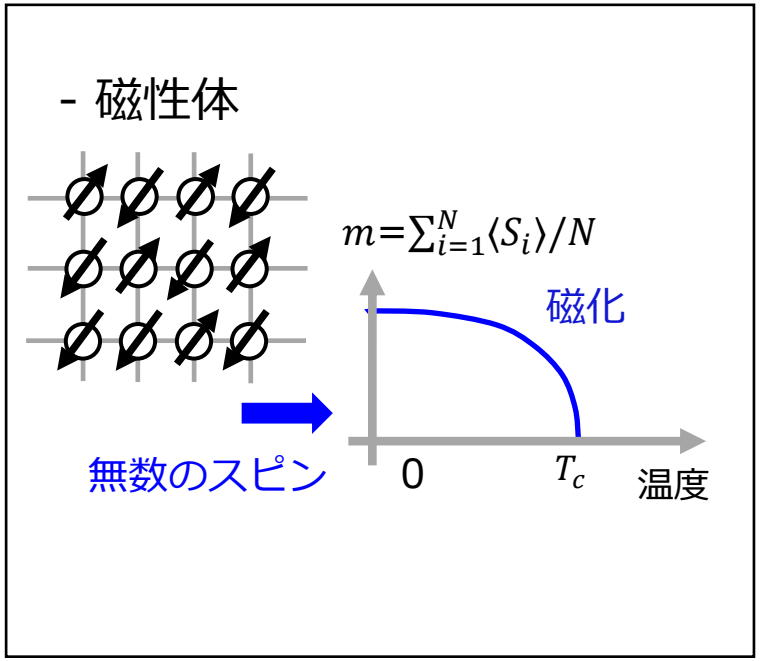
    [Karakida & Akaho, *"Learning Curves for Continual Learning in Neural Networks: Self-Knowledge Transfer and Forgetting"*, ICLR '22] + extensions

- 議論 (1枚)
  - 最適輸送と知識転移

力学・統計力学的解析

大自由度極限におけるニューラルネットの普遍的な挙動を理解・活用

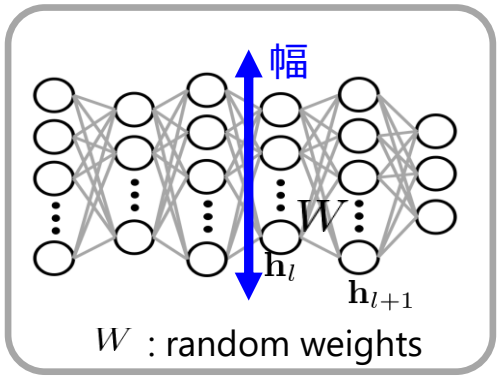
平均場, レプリカ, ランダム行列理論, …



[Amari '70-, Sompolinsky+ '88-]

ミクロ　　マクロ

$\mathbf{h}_{l+1}=\phi(W\mathbf{h}_l)$ ➡ $m_{l+1}=F(m_l)$



アーキテクチャ
ReLU, ResNet, Transformer, …
Batch Norm, Dropout

アルゴリズム
Backpropagation, SGD, SAM, Mixup, 蒸留 …

学習の目的・枠組み
教師あり/なし, 敵対的, 強化, 転移, 継続, …

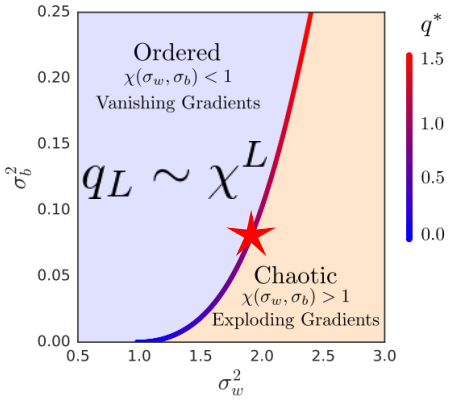
無限幅深層モデルの訓練性

- 訓練性を支配する行列

Jacobian of Backprop. : $\nabla_x f = \prod_{l=1}^{L} W_l^\top \phi'_{l-1}$

Fisher情報行列: $\nabla_\theta f^\top \ \nabla_\theta f$

Neural Tangent Kernel: $\nabla_\theta f \ \nabla_\theta f^\top$

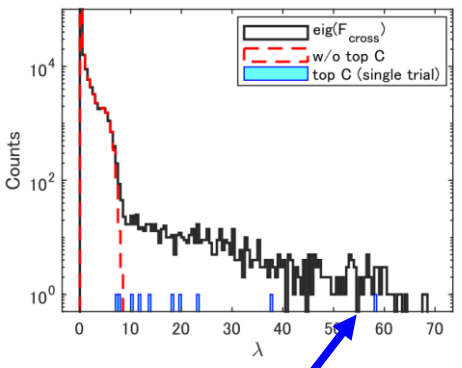


勾配消失-発散
⇔ 秩序-カオス相転移
[Poole+ '16], [Schoenholz+ '17]

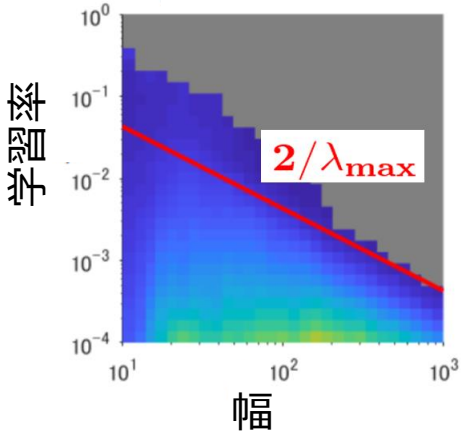
- 固有値の統計性

Fisherの場合

平均場 [K, Akaho & Amari, NeurIPS '19, NC '21]
ランダム行列理論 [Hayase & K, AISTATS '21]

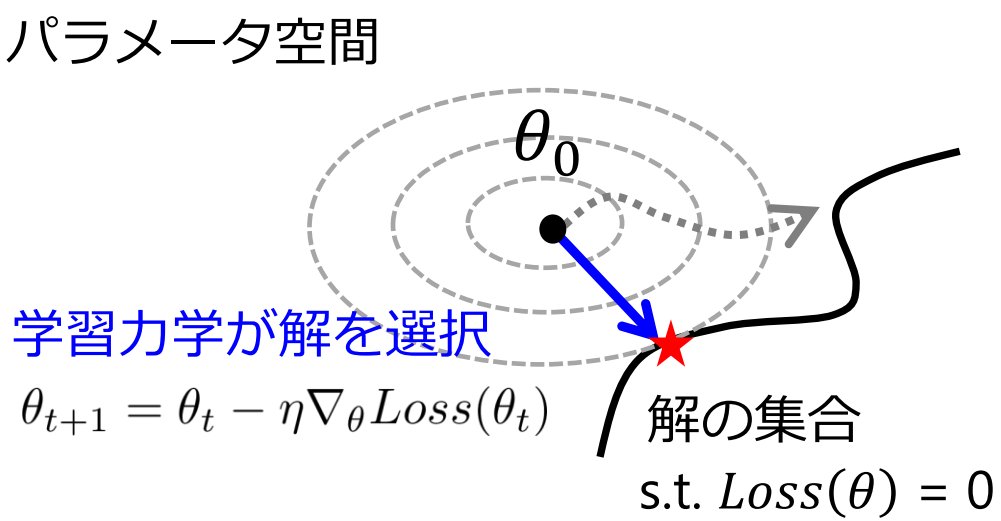


外れ値としての最大固有値λ_{max}



アルゴリズムの陰的バイアス

- 深層学習を根幹とした大規模モデルは過剰パラメータ系
  - **パラメータ数P > (≫) 訓練サンプル数N**
  - 大域解は無数に存在 $\{\theta^* | Loss(\theta^*) \simeq 0\}$



力学が可解な/特徴づけできる代表例

- 線形ネット

  $W_L \cdots W_2 W_1 x$

- 線形化 (NTK)

  $\nabla_\theta f(\theta_0)(\theta - \theta_0)$

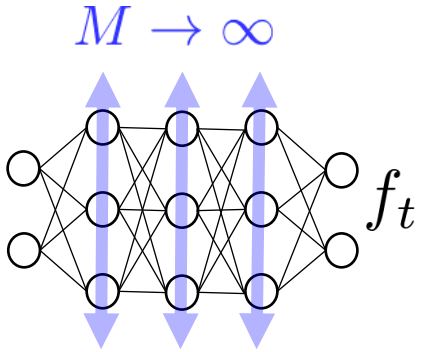
- Mean-field regime

  $\int d\mu(w_1, w_2) w_2 \phi(w_1 x)$

- (Langevin系)

解ける例1: Neural Tangent Kernel (NTK) Regime

[Jacot, Gabriel & Hongler, NeurIPS '18]



- 全結合モデル

$$u_l = \frac{\sigma_w}{\sqrt{M_{l-1}}} W_l h_{l-1} + \sigma_b b_l, \quad h_l = \phi(u_l)$$

- ガウス初期値　$W_{l,ij}, b_{l,i} \sim \mathcal{N}(0,1)$

係数に注意 (NTK parameterization). 通常の初期値 $W \sim \mathcal{N}(0, 1/M)$ で学習率を$1/M$倍することに対応.

$$\frac{\partial u}{\partial W} = \frac{\sigma_w}{\sqrt{M}} h$$

- 訓練サンプル数N, 層数L, クラス数Cは幅Mに対して$O(1)$

- 平均二乗誤差ロスの最急勾配　$\theta_{t+1} = \theta_t - \eta \nabla_\theta \mathcal{L}(\theta_t)$

解ける例1: NTK Regime

初期値におけるNTKを $\Theta(x', x) = \nabla_\theta f_0(x') \nabla_\theta f_0(x)^\top$ とおく.

結果 [Jacot+ NeurIPS '18] [Lee+ NeurIPS '19]

$\eta < 2/\lambda_{max}(\Theta)$ のとき, 幅無限大の極限で, 任意のtとx'に対し,

$$f_t(x') = \Theta(x', x)\Theta^{-1}(I - (I - \eta\Theta)^t)(y - f_0) + f_0(x')$$

- **線形モデルの訓練と等価**

$$\frac{d\theta_t}{dt} = \eta \nabla_\theta f_t^\top (y - f_t) \quad \Rightarrow \quad \frac{df_t}{dt} = \eta \underline{\nabla_\theta f_t \nabla_\theta f_t^\top}(y - f_t)$$

$$=: \Theta_t$$

$$\frac{df_t^{\mathrm{lin}}}{dt} = \eta \underline{\Theta_0}(y - f_t^{\mathrm{lin}})$$

$$f_t^{\mathrm{lin}}(x) = f_0(x) + \underbrace{\nabla_\theta f_0(x)^\top}_{\sim M^{-1/2}} \underline{(\theta_t - \theta_0)}$$

微小変化するパラメータがたくさんあるので
出力は $\mathcal{O}(1)$ で変化

$$\eta \nabla_\theta f \sim M^{-1} \cdot M^{-1/2}$$

$$P \cdot M^{-2} \sim 1$$

解ける例1: NTK Regime

$$f_t(x') = \Theta(x', x)\Theta^{-1}(I - (I - \eta\Theta)^t)(y - f_0) + f_0(x')$$

- 訓練済みモデルは**Kernel ridge-less regression (KRR)**と等価.

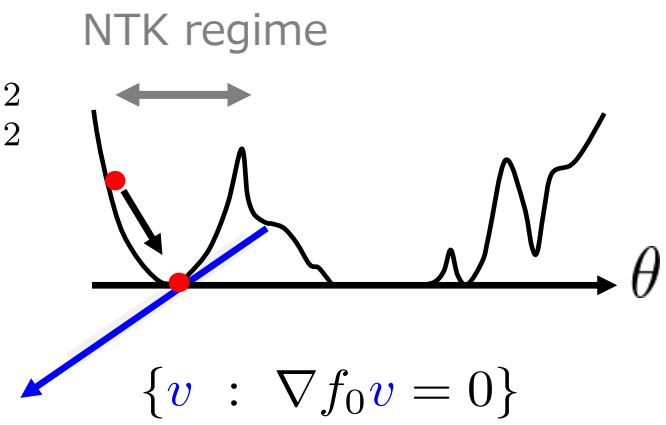
$$\langle f_\infty(x') \rangle_{\text{ini.}} = \Theta(x', x)\Theta(x, x)^{-1} y$$

- L2ノルム最小解として特徴づけられる

$$\Delta\theta = \theta - \theta_0,\ E(\Delta\theta) = \|y - (f_0 + \nabla f_0 \Delta\theta)\|_2^2$$

とおくと,

$$\text{argmin}_{\Delta\theta} \|\Delta\theta\|_2 \ s.t.\ E(\Delta\theta) = 0$$



近似自然勾配法の解析

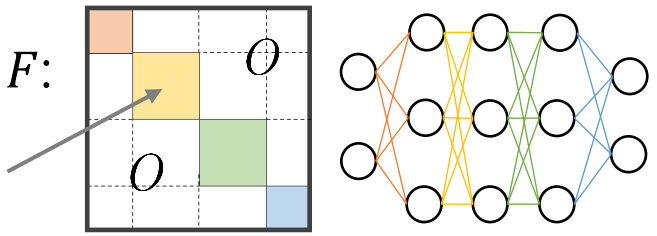
[Karakida & Osawa (東工大 → DeepMind), NeurIPS 2020]

$$\theta_{t+1} = \theta_t - \eta F_t^{-1} \nabla_\theta \mathcal{L}(\theta_t)$$

- Fisher情報行列 ($= F$)の計算コストを下げるため, 近似が使われる
  層ブロック対角, **K-FAC**, **ユニット対角** …

ヒューリスティックな置き換え

$$E[A_l \otimes B_l] \sim E[A_l] \otimes E[B_l]$$



- NTK regimeでは, 適切な学習率とdampingのもと, これら近似自然勾配法の訓練ダイナミクスf_tは**最速** (= exactと一致)

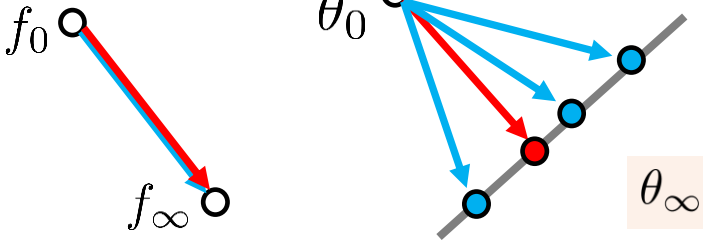
$$\nabla_\theta f (F + \rho I)^{-1} \nabla_\theta f^\top = \alpha I \quad \textbf{：等方性条件}$$

$$f_t = y + (I - \eta \Theta)^{\alpha t} (f_0 - y)$$

$$\nabla f \left((A + \rho I) \otimes (B + \rho I)\right)^{-1} \nabla f^\top$$
$$= \delta\delta^\top \left(\delta\delta^\top + \rho I\right)^{-1} \odot hh^\top \left(hh^\top + \rho I\right)^{-1}$$
$$\to I \quad (\rho \to 0)$$

近似自然勾配法の解析

- なお, 関数空間でのダイナミクスは近似手法間で一致するが, パラメータ空間では一致しない



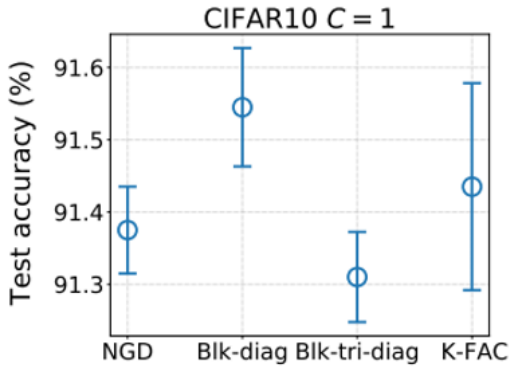
$$\theta_{\infty,\text{approx.}} = \alpha^{-1} G_0^{-1} J_0^\top y$$

$$\theta_{\infty,\text{ original NGD}} \left(= \theta_{\infty,\text{GD}}\right)$$

- 近似ごとに選ばれる解は異なり, 汎化性能も異なりうる

$$\text{argmin}_\theta \frac{1}{2N} \|y - J_0\theta\|_2^2 + \frac{\lambda}{2}\theta^\top G_0 \theta$$

in the ridge-less limit ($\lambda \to 0$)



解ける例2: 対角線形ネットにおける解選択

$$f(x) = \sum_{i=1}^{D} \underbrace{(w_{+,i}^2 - w_{-,i}^2)}_{=:\beta_i} x_i$$

- 初期値パラメータスケールα [Woodworth, …, Srebro, COLT '20]

勾配流: $dw/dt = -\nabla_w \mathcal{L}$

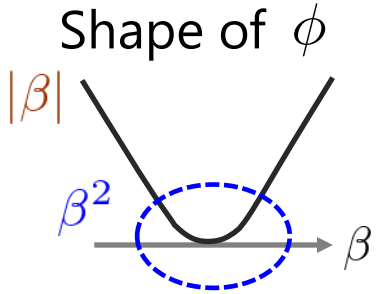
大域解の形: $\beta^{\infty}(\alpha) = \underset{\beta \in \mathbb{R}^D \text{ s.t. } X\beta = y}{\arg\min} \phi_\alpha(\beta)$

(i) $\alpha \gg 1$: NTK (Lazy) regime

$\phi_\alpha(\beta) \sim \|\beta\|_2^2$ L2ノルム正則化

(ii) $\alpha \ll 1$: Rich regime

$\phi_\alpha(\beta) \sim \|\beta\|_1$ L1ノルム正則化



解ける例2: 対角線形ネットにおける解選択

- 初期値パラメータスケールα [Woodworth, …, Srebro, COLT '20]

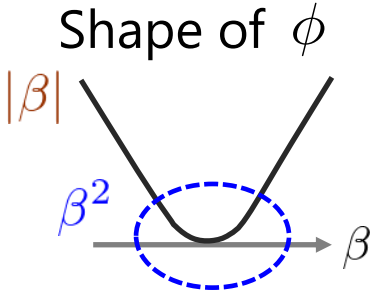
大域解の形: $\beta^{\infty}(\alpha) = \underset{\beta \in \mathbb{R}^D \text{ s.t. } X\beta = y}{\arg\min} \phi_\alpha(\beta)$

(i) $\alpha \gg 1$: NTK (Lazy) regime

$\phi_\alpha(\beta) \sim \|\beta\|_2^2$ L2ノルム正則化

(ii) $\alpha \ll 1$: Rich regime

$\phi_\alpha(\beta) \sim \|\beta\|_1$ L1ノルム正則化



- アルゴリズム "実効的な"αを下げる. Rich regimeへのバイアス
  - SGDノイズの大きさ [Pesme+, '21]
  - 離散更新ステップの大きさ [Nacson+, '22]
  - 勾配正則化 [RK, Takase, Hayase & Osawa, arXiv 2210.02720] $\mathcal{L}(\theta) + \lambda\|\nabla\mathcal{L}(\theta)\|^2$

Outline

- 研究背景
  - ニューラルネットの統計力学的解析
  - アルゴリズムと解選択
    NTK, 対角線形ネット

- 継続学習における知識転移
  - 汎化誤差のレプリカ解析
  - 自己知識転移と忘却

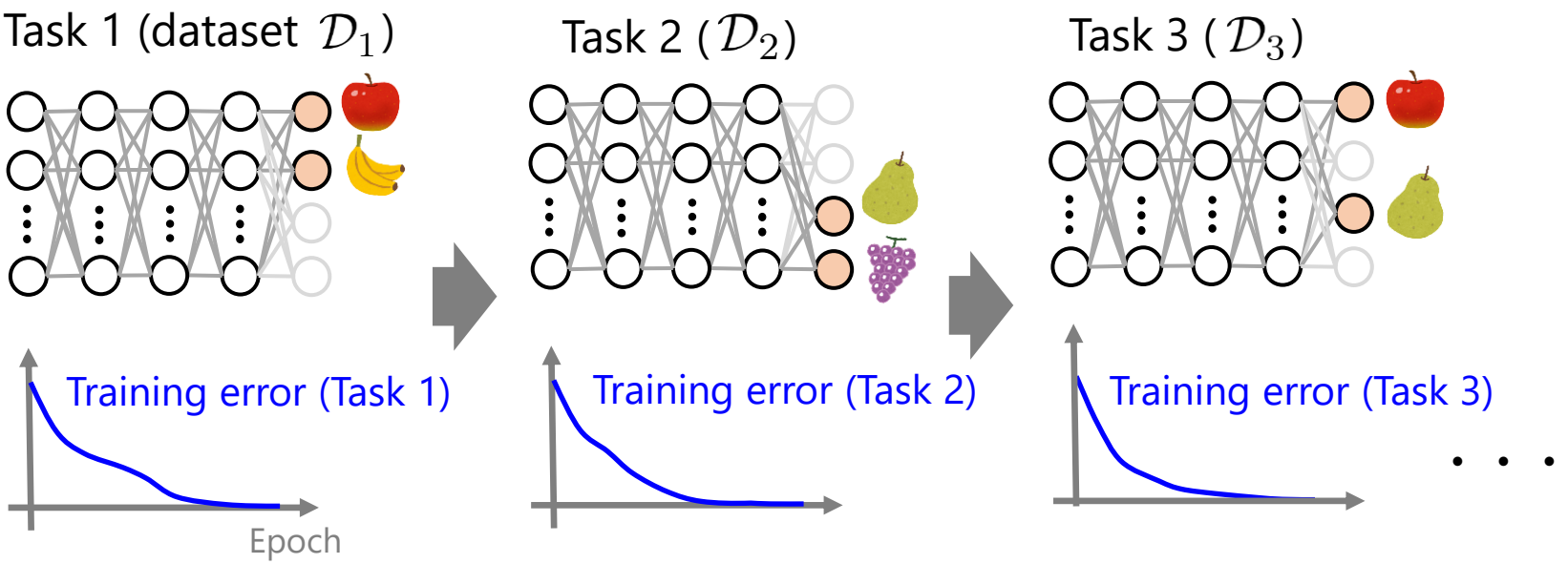
  [Karakida & Akaho, *"Learning Curves for Continual Learning in Neural Networks: Self-Knowledge Transfer and Forgetting"*, ICLR '22] + extensions

- 議論
  - 最適輸送と知識転移

継続学習 (Continual Learning)

a.k.a. Lifelong learning

同じモデルを異なるタスク間で訓練しつづける
- "タスク"はデータセットぐらいの意味
- 前タスクのサンプルは保存しない (メモリ制限, privacyなど)



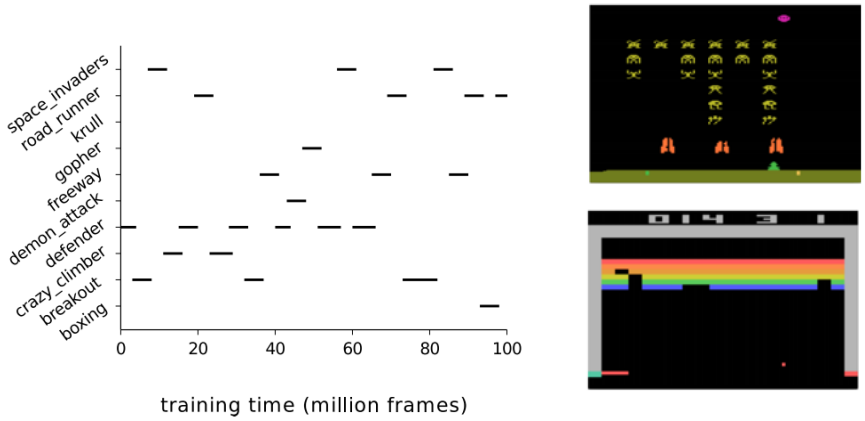
"ダイナミクスのダイナミクス"

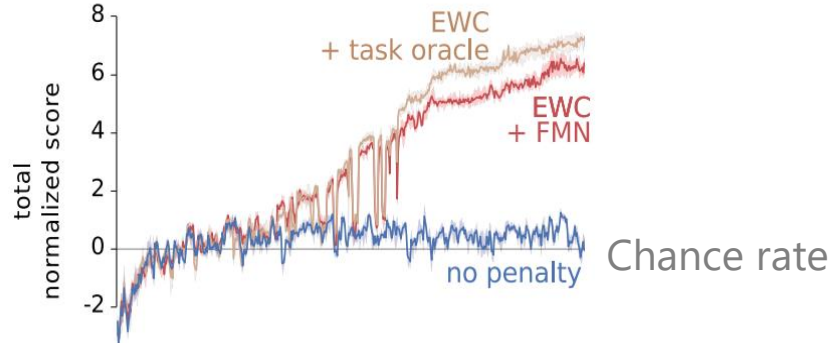
継続学習 (Continual Learning)

- 強化学習

Deep Q-Network on Atariゲーム

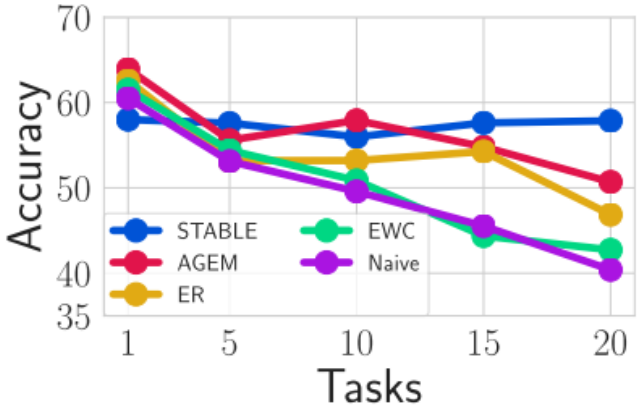
[Kirkpatrick+ PNAS '17]





- 教師あり学習

- 追加学習の例
ResNet on クラス分割CIFAR-100.
(タスク間でクラスは非共通)



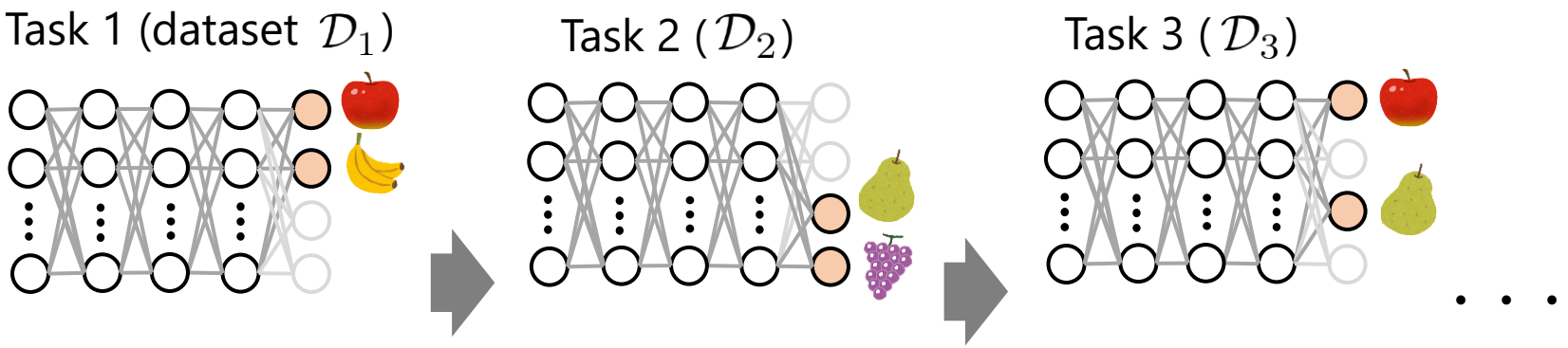
[Mirzadeh+ ICLR '20]

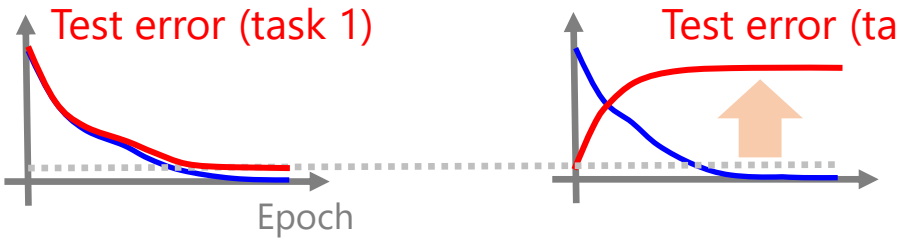
継続学習 (Continual Learning)

同じモデルを異なるタスク間で訓練しつづける

- すべてのタスクで高い予測性能を達成したい

  転移学習との違い





破滅的忘却
(Catastrophic Forgetting)

[McCloskey & Cohen (1989)]

NTKレジームにおける継続学習

[Bennani, Doan & Sugiyama '20] [Doan, Bennani, Mazoure, Rabusseau & Alquier '21]

タスクnまで訓練を終えたモデル(f_n)のパラメータをθ_nとしたとき,
$\mathcal{L}_{n+1}(\theta)$ 最適化の初期値が θ_n

$$\begin{aligned}\theta(0) &\leftarrow \theta_n \\ \theta(t+1) &\leftarrow \theta(t) - \eta \nabla_\theta \mathcal{L}_{n+1}(\theta(t)) \\ \theta_{n+1} &\leftarrow \theta(\infty)\end{aligned}$$

$f_{n+1}(\theta) = f_n + \nabla_\theta f_0^\top (\theta - \theta_n)$ でタスク(n+1)を学習.

継続学習

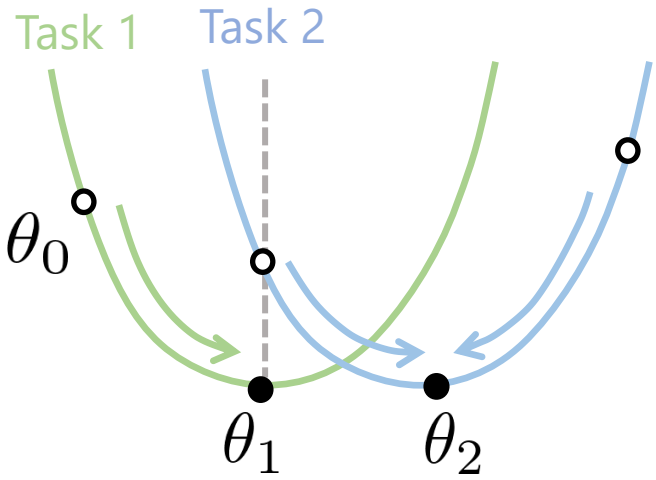
$$f_n(x') = f_{n-1}(x') + \Theta(x', X_n)\,\Theta(X_n)^{-1}(y_n - f_{n-1}(X_n))$$
$$\theta_n - \theta_{n-1} = \nabla_\theta f_0(X_n)^\top \Theta(X_n)^{-1}(y_n - f_{n-1}(X_n))$$

タスクnの訓練サンプル: $\{X_n, y_n\}$　NTK行列: $\Theta(x', x) = \nabla_\theta f_0(x') \nabla_\theta f_0(x)^\top$

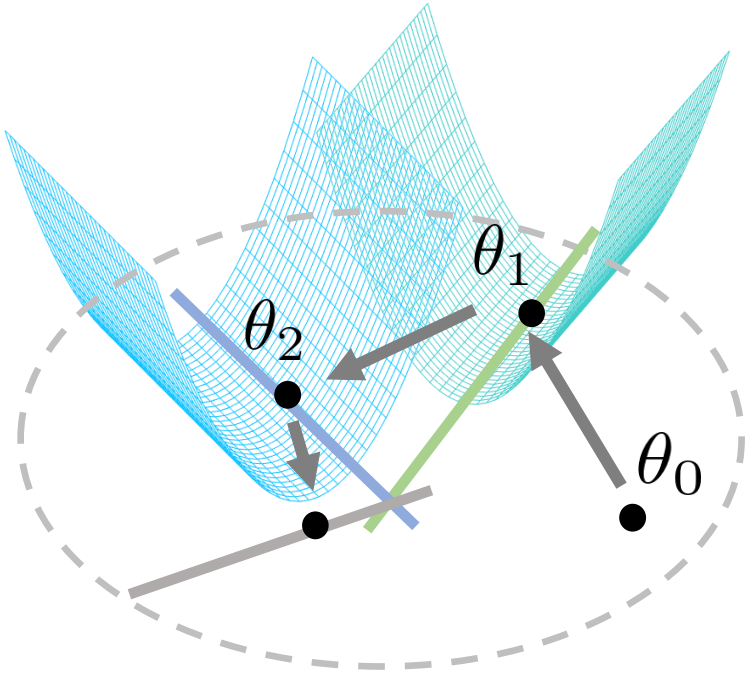
- Rademacher complexityによる汎化バウンド [Bennani+ '20]

線形化モデルの継続学習

- Under-parameterization



- Over-parameterization



Over-parameterizationでは解が各タスクの初期値に依存.

$$\theta_n = \mathrm{argmin}_\theta \|\theta - \theta_{n-1}\|_2 \ s.t.\ y = f_{n-1} + \nabla_\theta f_0^\top (\theta - \theta_{n-1})$$

Kernel Ridge(-less) Regression; KRR

$$f^* = \arg\min_{f\in\mathcal{H}} \frac{1}{2\lambda}\sum_{\mu=1}^{N}\left(f\left(\mathbf{x}^\mu\right)-y^\mu\right)^2 + \frac{1}{2}\langle f, f\rangle_{\mathcal{H}} \qquad \mathcal{H}\text{: RKHS}$$

Mercer分解: $\int d\mathbf{x}' p\left(\mathbf{x}'\right)\Theta\left(\mathbf{x},\mathbf{x}'\right)\phi_i\left(\mathbf{x}'\right) = \eta_i\phi_i(\mathbf{x}) \quad (i=1,2,\ldots)$

固有値

基底: $\psi_i \equiv \sqrt{\eta_i}\phi_i$

解: $f^*(\mathbf{x}) = \sum_{i=1}^{\infty} \mathbf{w}_i^*\psi_i(\mathbf{x}), \quad \mathbf{w}^* = \left(\mathbf{\Psi}\mathbf{\Psi}^\top + \lambda\mathbf{I}\right)^{-1}\mathbf{\Psi}\mathbf{y}$

- 一般にNTKは内積カーネル $\Theta\left(\mathbf{x}',\mathbf{x}\right) = f\left(\mathbf{x}'^\top\mathbf{x}\right)$
- 入力は超球面上($\|\mathbf{x}\|^2 = 1$) の一様分布を仮定
- **教師信号 (Target)** を仮定

$$y^\mu = \bar{f}(\mathbf{x}^\mu) + \epsilon^\mu \qquad \epsilon^\mu \sim \mathcal{N}(0,\sigma^2),\ \sigma^2 \geq 0$$

Target $\bar{f}$は二乗可積分. ここでは, 生徒のRKHSに含まれるとする

$$\bar{f}(x) = \sum_i \bar{w}_i\sqrt{\eta_i}\phi_i(x), \quad \sum_i \bar{w}_i^2 < \infty$$

KRRのレプリカ解析

[Bordelon, Canatar & Pehlevan, ICML '20] …

- 汎化誤差E_gを**典型評価 (平均評価)**

$$E_g = \left\langle \int d\mathbf{x}\, p(\mathbf{x}) \left(f^*(\mathbf{x}) - \bar{f}(\mathbf{x})\right)^2 \right\rangle_{\mathcal{D}}$$

訓練サンプルセットに対する平均
(配位平均に対応)

訓練サンプル数Nが十分に大きいとき,
レプリカ法 (*w/* レプリカ対称性) より,

$$E_g = \frac{1}{1-\gamma} \sum_{i=0}^{\infty} \eta_i \bar{w}_i^2 \left(\frac{\kappa}{\kappa + N\eta_i}\right)^2 + \frac{\gamma}{1-\gamma}\sigma^2$$

η_i ：カーネル固有値
$\bar{w}_i^2$ ：Target係数
σ^2 ：Targetノイズ
λ ：リッジ正則化

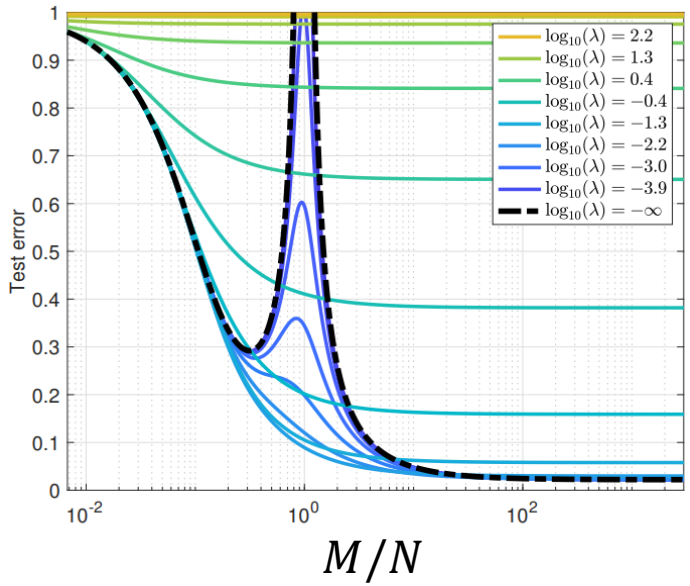
- 固有値モードによる展開. 二重降下・多重降下現象の相図が書ける.
- 固有値から再帰的に定まる定数 κ

$$\kappa = \lambda + \sum_i \frac{\kappa\eta_i}{\kappa + N\eta_i}\ , \qquad \text{定数}\ \gamma = \sum_{i=0}^{\infty} \frac{N\eta_i^2}{(\kappa + N\eta_i)^2}$$

過剰パラメータ系の学習曲線

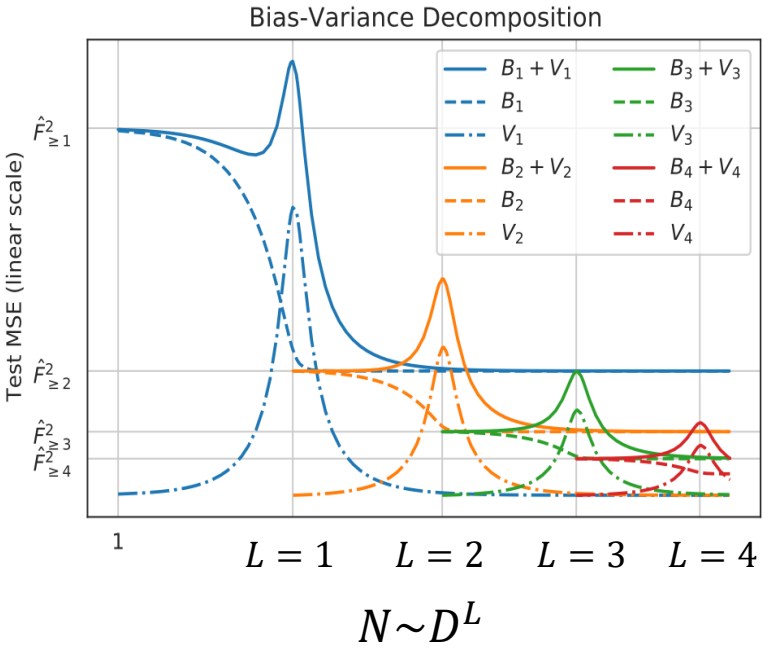
$$f(x) = \theta^\top \phi(JX), \quad J_{ij} \sim \mathcal{N}(0, 1/M)$$

Random feature Regression



[Mei & Montanari, '19]

Kernel Ridge(-less)

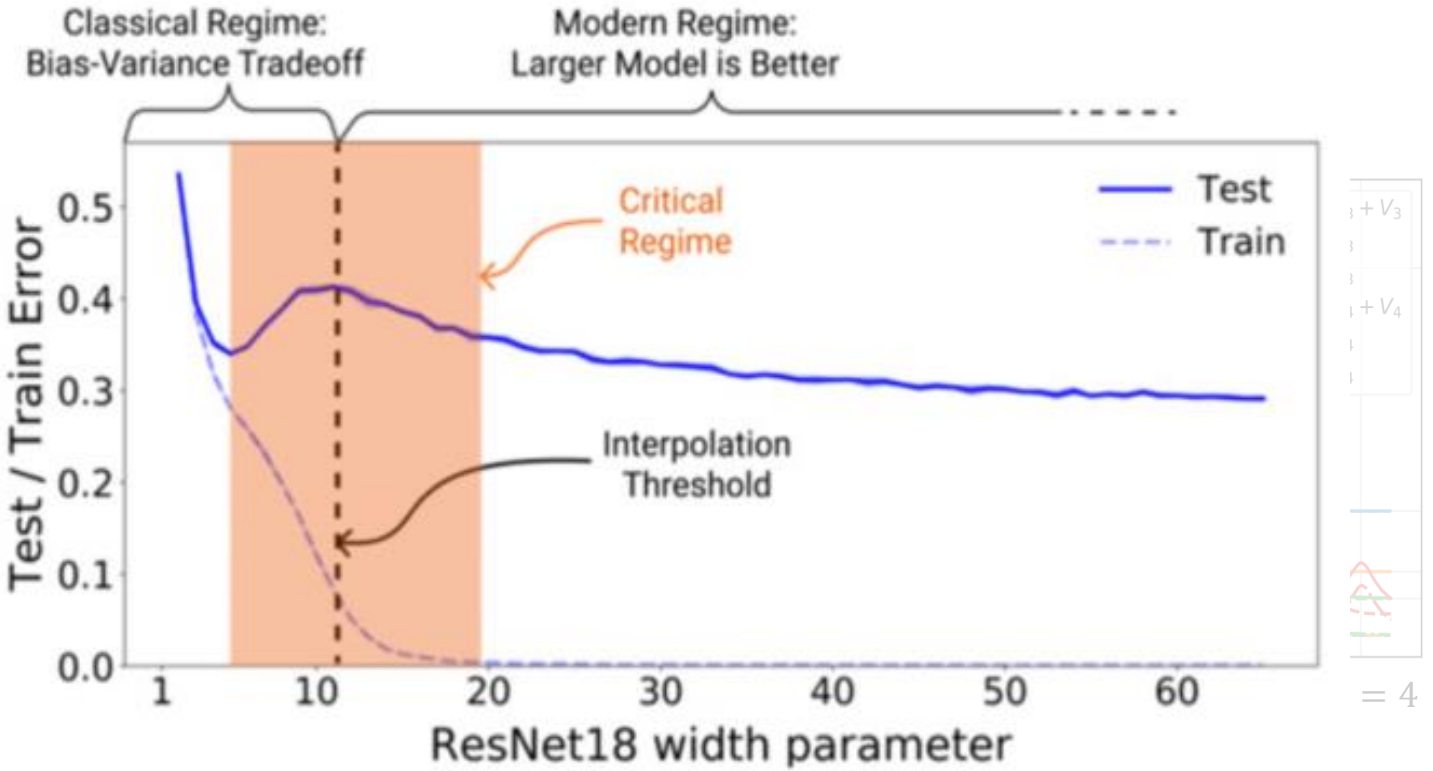


[Xiao & Pennington, '22]

M: Width N: Sample size, D: Input dim.

過剰パラメータ系の学習曲線

$$f(x) = \theta^\top \phi(JX), \quad J_{ij} \sim \mathcal{N}(0, 1/M)$$



[Nakkiran+, '20]

設定: 2タスク間の知識転移

- 相関した教師ターゲットを2つ用意. **Target Similarity (ρ)**を導入.

$$\bar{f}_A(x) = \sum_i \bar{w}_{A,i}\psi_i(x),\ \bar{f}_B(x) = \sum_i \bar{w}_{B,i}\psi_i(x)$$

$$\rho = \langle \bar{f}_A, \bar{f}_B \rangle_{\mathcal{H}} / (\|\bar{f}_A\|_{\mathcal{H}} \|\bar{f}_B\|_{\mathcal{H}})$$

- 教師係数をガウス乱数で生成 $[\bar{w}_{A,i}, \bar{w}_{B,i}] \sim \mathcal{N}(0, \eta_i \begin{bmatrix} 1 & \rho \\ \rho & 1 \end{bmatrix})$

- 訓練サンプルの生成
  - 入力サンプルは, 超球面上一様に i.i.d.生成 *Domain shift無し

$$x_A^\mu, x_B^\mu \overset{\text{i.i.d.}}{\sim} p(x)\ ,\ \ \|x_A^\mu\| = \|x_B^\mu\| = 1$$

$$y_A^\mu = \bar{f}_A(x_A^\mu) + \varepsilon_A^\mu \quad (\mu = 1, ..., N_A)$$
$$y_B^\mu = \bar{f}_B(x_B^\mu) + \varepsilon_B^\mu \quad (\mu = 1, ..., N_B)$$

設定: 2タスク間の知識転移

- 汎化誤差の定義　$f_{A\to B}$: タスクAからBまで継続学習したモデル

$$E_{A\to B}(\rho) = \left\langle \int dx p(x) \left(\bar{f}_B(x) - f_{A\to B}(x)\right)^2 \right\rangle_{\{\mathcal{D}_A, \mathcal{D}_B\}}$$

後続タスクでの予測

$$E^{back}_{A\to B}(\rho) = \left\langle \int dx p(x) \left(\bar{f}_A(x) - f_{A\to B}(x)\right)^2 \right\rangle_{\{D_A, D_B\}}$$

先行タスクでの予測
(Backward Transfer)

結果: 2タスク間のKnowledge Transfer

訓練サンプル数N_A, N_Bが十分に大きいとき, レプリカ法のもと,

$$E_{A\to B}(\rho) = \sum_i^{\infty} \left[2(1-\rho)\left(1-q_{A,i}\right) + \frac{q_{A,i}^2}{1-\gamma_A} \right] E_{B,i}$$

$$E_{A\to B}^{back}(\rho) = \sum_i^{\infty} \left[2(1-\rho)\left(1+F_{\gamma_B}\left(q_{A,i}, q_{B,i}\right)\right)\eta_i^2 + \frac{q_{A,i}^2}{1-\gamma_A} E_{B,i} \right]$$

$E_{B,i} = q_{B,i}^2 \eta_i^2/(1-\gamma_B)$ ：単一タスクの汎化誤差に対応

定数κ, γは, 単一タスクのときと同様に計算

$$q_{A,i} = \kappa_A/(\kappa_A + N_A\eta_i),\ q_{B,i} = \kappa_B/(\kappa_B + N_B\eta_i),\ F_{\gamma_B}(a,b) = b(a-2) + b^2(1-a)/(1-\gamma_B)$$

導出の方針: $E_g = \frac{1}{1-\gamma}\sum_{i=0}^{\infty} \eta_i \bar{w}_i^2 \left(\frac{\kappa}{\kappa + N\eta_i}\right)^2$

教師係数で
展開できる点に注目.

$$\left\langle \int dx p(x)(\bar{f}(x) - f_{n+1}(x))^2 \right\rangle_{\mathcal{D}} = \left\langle \int dx p(x)((w_{n+1}^* - (\bar{w} - \sum_{k=1}^{n} w_k^*))^\top \psi(x))^2 \right\rangle_{\mathcal{D}}$$

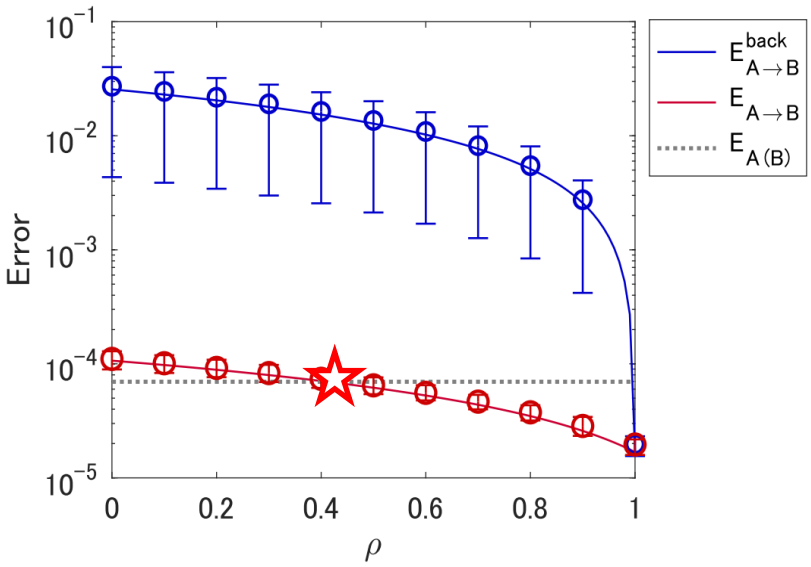
結果: 2タスク間の知識転移

- Critical task similarity

$$E_{A\to B}(\rho)/E_B \sim 2(1-\rho) \quad \textit{for} \quad N_A \gg N_B$$

$\rho > 1/2$のとき, Task B単体での汎化誤差(E_B)より改善 ($E_{A\to B} < E_B$).

= Positive forward transfer



実線：理論

点：NTK regime (幅4,000, GD訓練)

ReLU, $L = 3$, $N_A = N_B = 100$

結果: 2タスク間の知識転移

- Critical task similarity

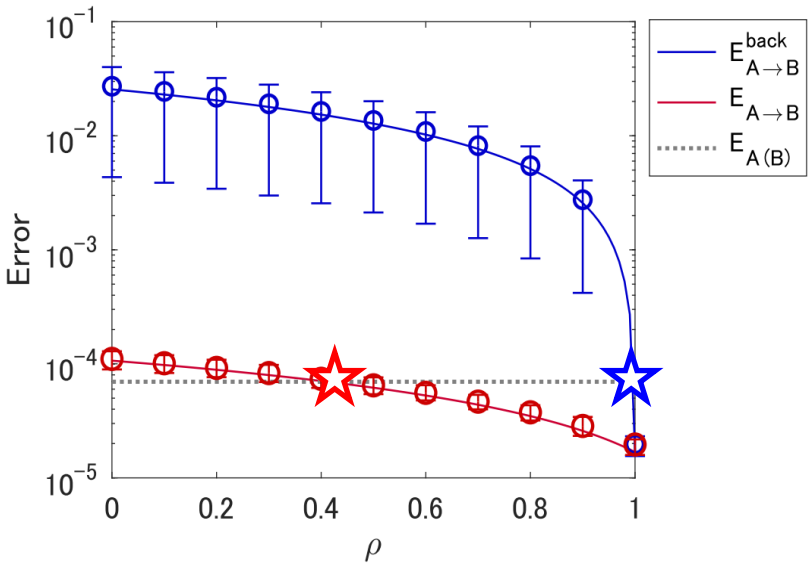
$$E_{A\to B}(\rho)/E_B \sim 2(1-\rho) \quad \textit{for} \quad N_A \gg N_B$$

$\rho > 1/2$のとき, Task B単体での汎化誤差(E_B)より改善 ($E_{A\to B} < E_B$).
= Positive forward transfer

- Negative backward transfer

$$E^{back}_{A\to B}(\rho) \sim \sum_i \eta_i^2 (1-\rho) \quad \textit{for} \quad N_A, N_B \gg 1$$

$\mathcal{O}(1)$ サンプル数に依存しない

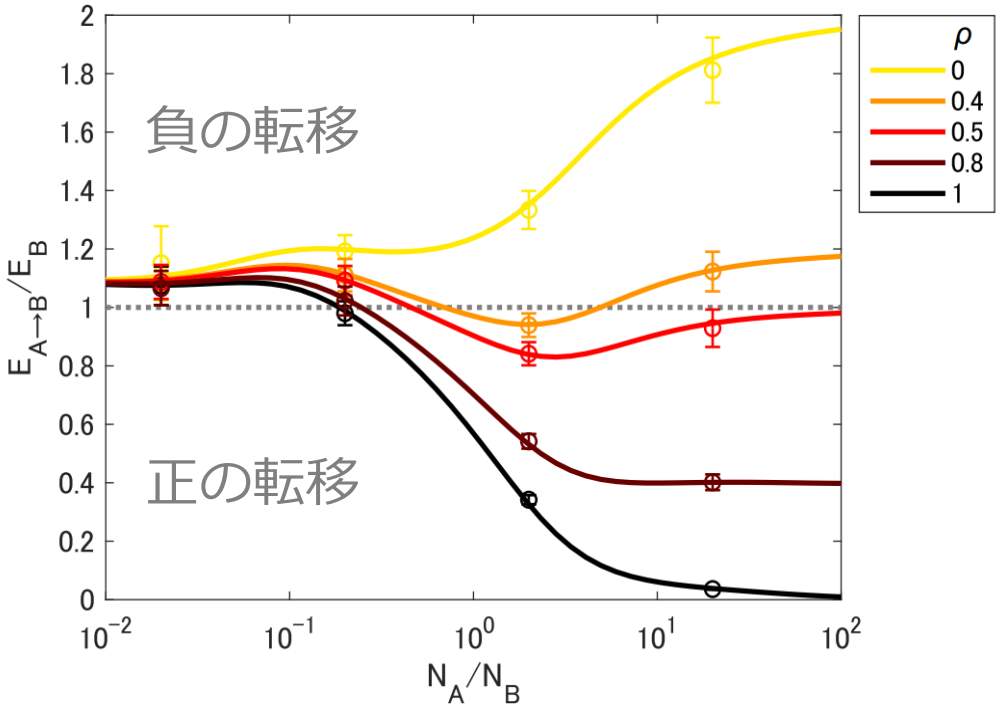


実線：理論
点：NTK regime (幅4,000, GD訓練)
ReLU, $L = 3$, $N_A = N_B = 100$

Task similarityがわずかに1より低いだけで負の転移 ($E^{back}_{A\to B} > E_A$).
破滅的忘却.

結果: 2タスク間の知識転移

$$E_{A\to B}(\rho) = \sum_{i}^{\infty} \left[2(1-\rho)\left(1-q_{A,i}\right) + \frac{q_{A,i}^2}{1-\gamma_A} \right] E_{B,i}$$



転移の成否は, タスク類似度とサンプル均衡 (N_A/N_B) に非単調に依存する

“自己”知識転移 (*self*-knowledge transfer)

設定: Task AとBのターゲットを同一とする $(\rho = 1)$

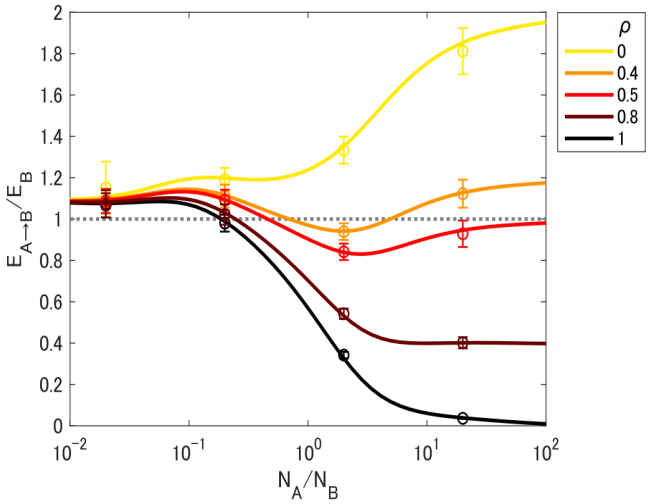
*このとき, $E_{A\to B}(1) = E^{back}_{A\to B}(1)$

- 訓練サンプル数が等しいとき $(N_A = N_B)$

$$E_{A\to B}(1) < E_{ave} < E_A = E_B$$

Positive transfer

E_{ave}：single-taskで訓練した
モデルのアンサンブル平均 $(f_A + f_B)/2$



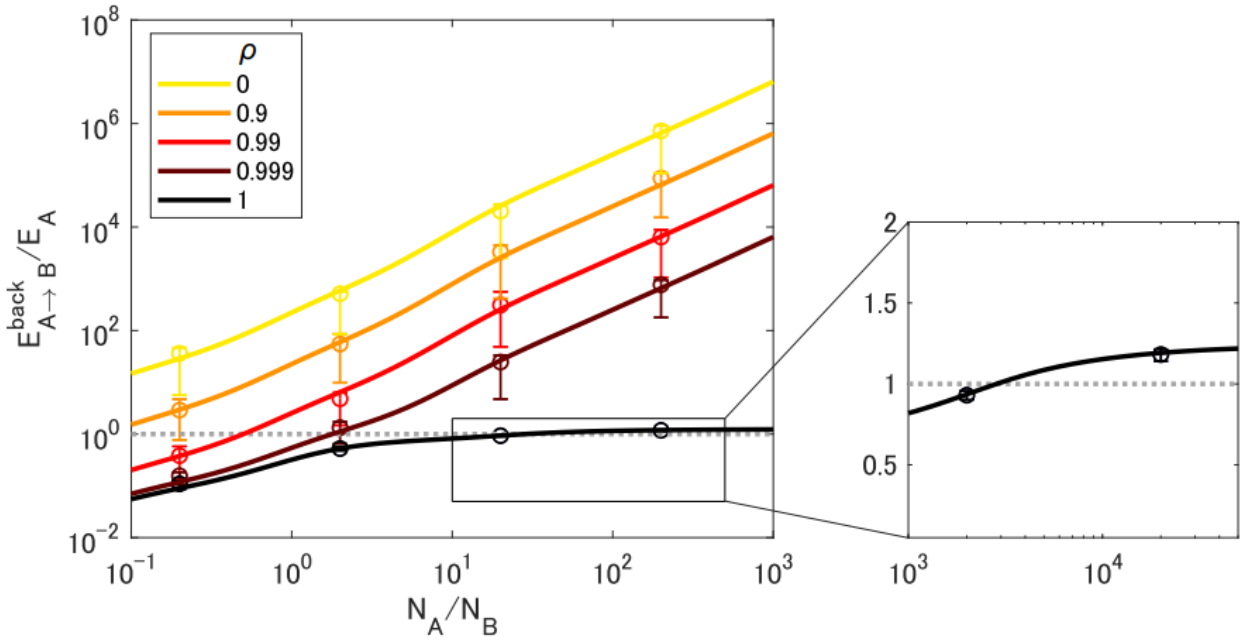
- 訓練サンプル数が不均衡のとき $(N_B \gg N_A)$

$$E_{A\to B}(1)/E_B \sim 1/(1-\gamma_A) > 1$$

Negative transfer

“自己”知識転移 (*self*-knowledge transfer)

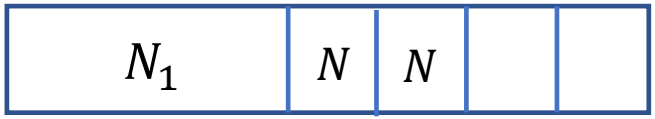
$$E^{back}_{A\to B}(1)/E_A \sim 1/(1-\gamma_B) > 1 \quad \textit{for} \quad N_A \gg N_B$$



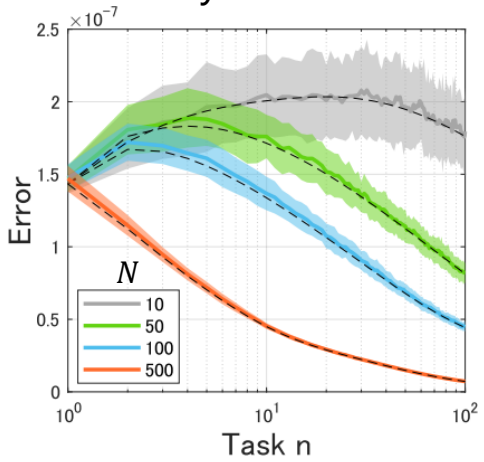
$N_A/N_B \gg 1$ (限られたサンプル数で事後に学習)のとき,
汎化性能が悪化. 自己知識忘却

多タスクの場合

- サンプル数の不均衡

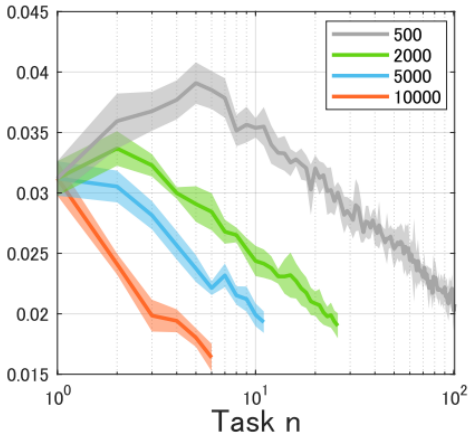


NTK regression
on synthetic data



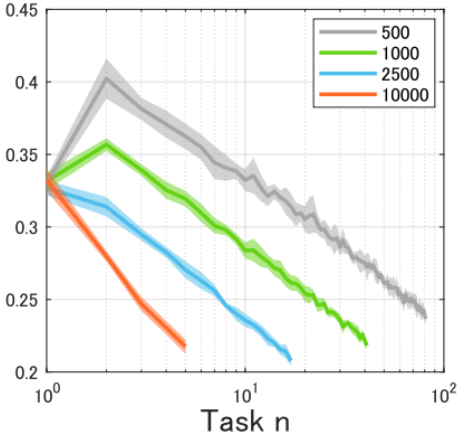
$N_1 = 4{,}000$

MLP on MNIST
(width 256, SGD)



$N_1 = 10^4$

ResNet-18 on CIFAR-10
(SGD)

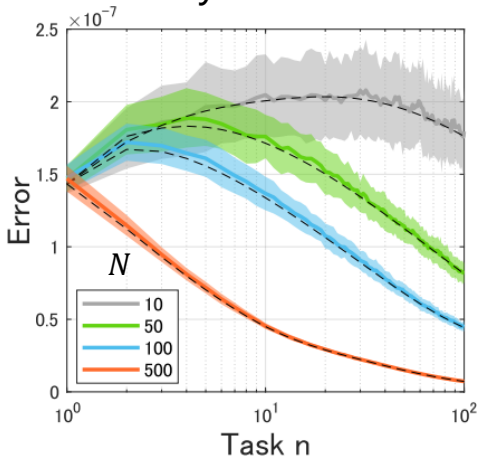


$N_1 = 10^4$

後続タスクのサンプル数(N) が小さいとき, 汎化誤差が増加
(自己知識忘却). 十分な数のタスクがあれば, 学習を続けることで減少.

多タスクの場合

NTK regression
on synthetic data



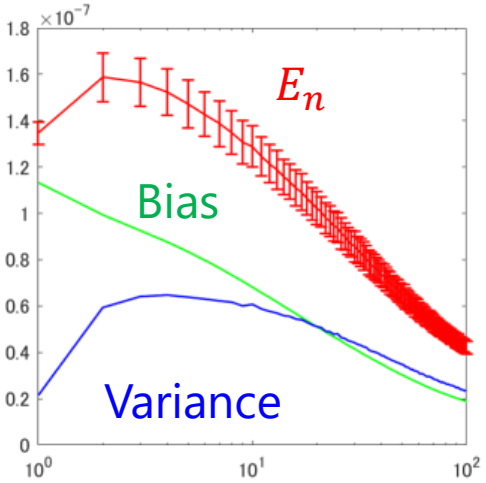
- 等分割 ($N_n = N$) ならば単調減少

$$E_{n+1} < E_n$$

- Bias項は不均衡でも常に単調減少

$$E_n = E_n^{(\mathrm{bias})} + E_n^{(\mathrm{variance})}$$

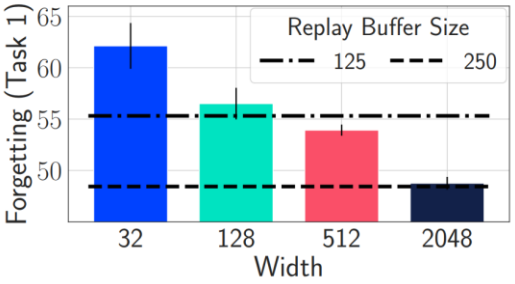
$$E_n^{(\mathrm{bias})} = \sum_i \bar{w}_i^2 \eta_i \prod_{m=1}^{n+1} q_{n,i}^2$$



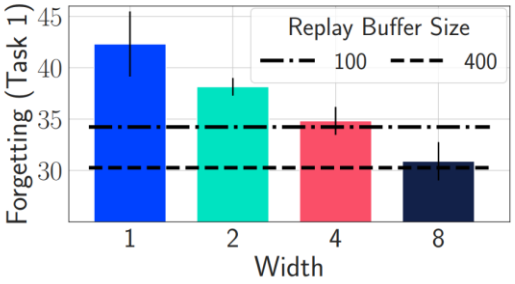
タスク数n

関連研究

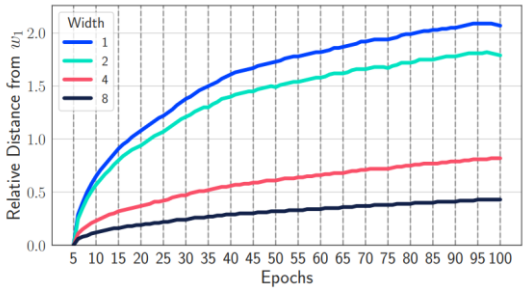
- NTK regime内なら, クラス追加学習による忘却は起こらない
  (parameter isolationの成立)



(a) MLP on Rotated MNIST



(b) WRN on Split CIFAR-100

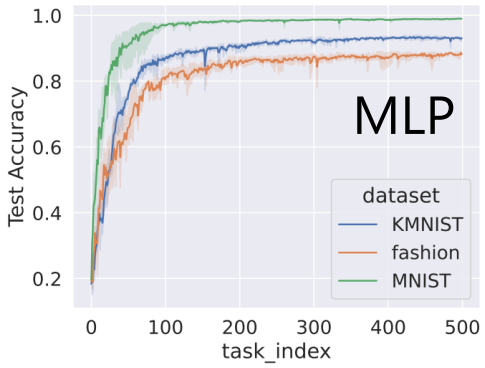


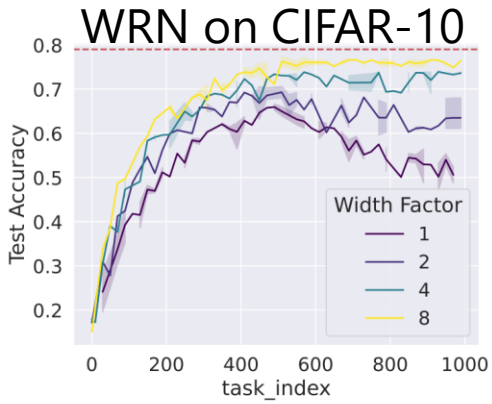
"Wide neural networks forget less catastrophically" [Mirzadeh+ ICML 22]

- サンプル数が均等なとき, 汎化誤差は減少

SCoLe (Scaling CL)

[Lesort+ arXiv2207.04543]





まとめ

可解モデルにおける力学/統計力学的解析の知見を紹介

深層学習で発展したアルゴリズムや学習の枠組みに隠された
法則や陰的バイアス

- 継続学習では自己知識忘却が現れる. タスク類似度だけでなく訓練サンプル数の均衡が重要.

課題

- 入力分布の一般化
  - **ドメインシフト** $(P_A(x) \neq P_B(x))$　線形回帰なら [Evron+ COLT '22]
  - "**Hidden manifold**"の導入
- 他の可解モデルクラス　線形ネット, Mean field regime (?)
  - 階層的なニューラルネット学習に特有な効果はあるか

  低類似タスクで忘却が軽減 [Lee, … , Goldt & Saxe, ICML '22]

終わりに: 知識転移とWasserstein距離

Optimal Transport Dataset Distance (OTDD)

[Alvarez-Melis & Fusi, NeurIPS '20]

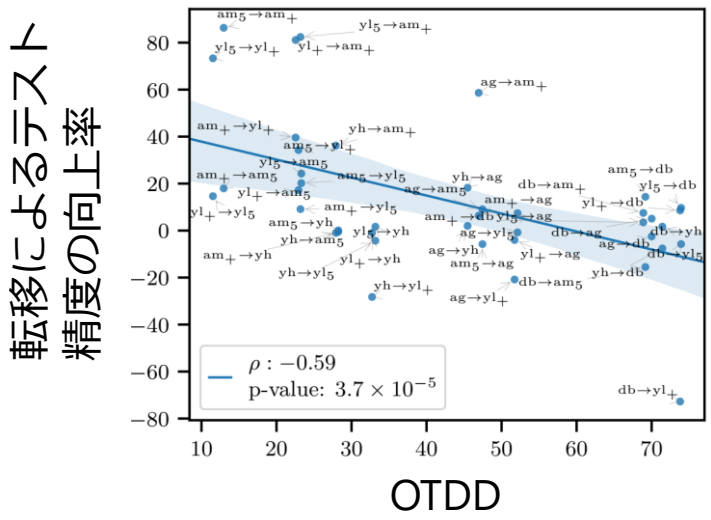
$$d_{\mathrm{OT}}\left(\mathcal{D}_A, \mathcal{D}_B\right) = \min_{\pi \in \Pi(\alpha,\beta)} \int_{\mathcal{Z}\times\mathcal{Z}} d_{\mathcal{Z}}\left(z, z'\right)^2 \pi\left(z, z'\right)$$

$$d_{\mathcal{Z}}\left(z, z'\right) = \left(d_{\mathcal{X}}\left(x, x'\right)^2 + d_{\mathcal{Y}}\left(y, y'\right)^2\right)^{1/2}$$

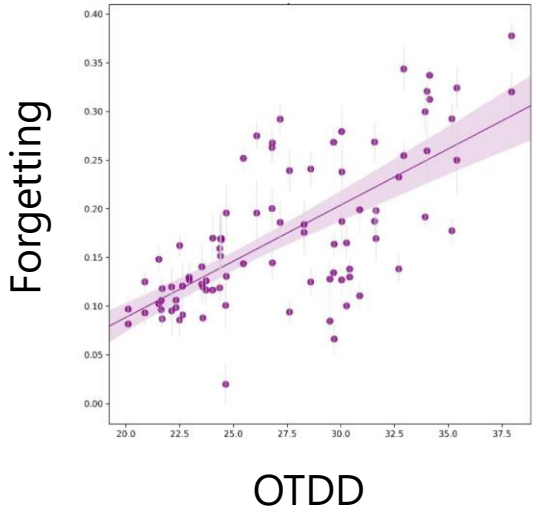
Sinkhorn距離

$P(x|y)$ で代用 & ガウスとみなして

Bures-Wasserstein距離



BERT on テキスト



ResNet34 on 分割CIFAR100 [Liu+arXiv:2208.11726]