参考資料1-1
科学技術・学術審議会
情報委員会（第47回）
令和8年5月12日

理研AIPセンター
汎用基盤技術研究グループ研究成果

グループディレクター　杉山 将

RIKEN AIP 革新知能統合研究センター
Center for Advanced Intelligence Project

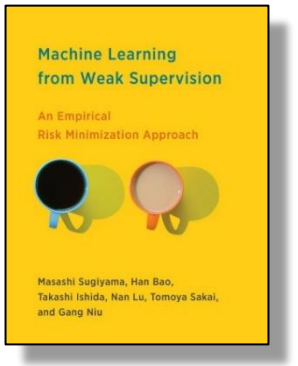
汎用Gの成果のまとめ(1ページの概要)2

- 目標:
  1. 最先端の機械学習技術の原理の解明
  2. 既存技術では対応できない難問を解決する,
     新たな機械学習技術の創出
  3. 日本の機械学習研究の国際的な認知度の向上
- 成果の概要の概要:
  1. 深層ニューラルネット, トランスフォーマー, 拡散モデルなどの高い予測・推論・生成能力を数学的に証明. スパース最適化, 多段階最適化, 超高次元最適化などに対する理論保証付き最適化法を開発.
  2. 不完全な教師情報だけからでも学習可能な技術, データから因果関係を推論できる技術, 不確定性を考慮した大規模適応学習技術を開発.
  3. 上記の成果が評価され, ICLR, AISTATS, ICCOPT等の主要国際会議で基調講演を実施. NeurIPS, ICML, ICLR, AISTATS, ISMP等の理事・運営委員長等に選出.
- 次頁以降, 各成果の概要を示します.

不完全情報学習（杉山TD）

- 既存の機械学習の弱点3つを克服する革新的な新技術を開発

1. 大量の教師データが必要：
   - 弱い教師情報からでも学習できる新技術を開発
   - MIT Pressより全貌をまとめた英語の専門書を出版
   - 主要論文の累計被引用数は4000+回



2. 正確な教師データが必要：
   - 雑音を含む教師情報からでも学習できる新技術を開発
   - NeurIPS’18発表論文は被引用3000+回（主要論文累計6800+回）

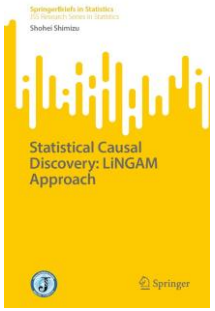
3. バイアスのない学習データが必要：
   - 15年前に開発した転移学習技術が生成AIの基盤技術としてリバイバル
   - 任意のバイアスに対応できる革新技術へと発展
   - 主要論文の累計被引用数は12100+回

- 機械学習3大国際会議の一つ
  ICLR2023にて日本人初の基調講演

因果推論（清水TD）

- AIをさらに発展させる因果推論の新技術を開発
  - AI for Scienceが盛り上がる今こそ、因果推論は不可欠

1. 従来 因果構造が既知である必要：
   - 潜在共通原因があってもデータ駆動で因果構造を探索する新技術を開発
   - Springerより一連の研究をまとめた英語の専門書を出版



2. 従来 科学者・実務家の専門知識が必要：
   - LLMと統計的因果構造探索の連携
   - 全国規模の保険者データベースによるPoCの実施
     - JST CREST (信頼されるAI領域)

3. 従来 科学者・実務家用ソフトウェアの不足：
   - Pythonパッケージの開発2つ: 月2万・10万ダウンロード以上
   - 商用ソフトウェア採用3件 (1件監修)

Causalas

- 因果推論専門国際WS・会議での招待講演・パネリスト
  - ワークショップ at UAI2023, KDD2021, NeuIPS2020
  - 会議 Pacific Causal Inference Conference 2024, 2020

適応学習（KhanTD）

- AI学習を持続可能にすべく，コストを劇的に下げたい：
  - 巨大データ・巨大計算機が不要な計算効率の良いアルゴリズム
  - 既存の学習済みモデルの継続的な更新・再利用
- 提案技術：ベイズ学習規則
  - 既存の主要な学習アルゴリズムを含む統合的な枠組み
  - 大規模言語モデルの学習を可能にした初のベイズ学習技術
  - NeurIPS2021で開催されたコンペで優勝
  - 変化する環境に適応できる継続学習へと拡張
  - JST-CRESTに採択（2.2億円）
- AI・ベイズ学習コミュニティにおける多数の招待講演：
  - NeurIPS2019にてチュートリアル（7000+人）
  - ISBA2026，BayesComp2025，EurIPS2025（2000+人），CoLLAs 2024にて招待講演
- 機械学習の主要国際会議にてプログラム委員長等を歴任：
  - ICLR 2024, AISTATS 2025, 2026

深層学習理論（鈴木TD）

- 最先端の機械学習技術を安心して使うための数学的な原理解明
  - 深層学習の高次元学習問題における優位性を証明：
    - 深層ニューラルネット：学習が難しい方向を見つけて集中的に学習
    - 拡散モデル：データの低次元構造を抽出して効率良く学習
    - トランスフォーマー：重要なトークンを選ぶことにより長文を扱える
  - 困難な深層モデルの最適化の原理の解明：
    - 確率的勾配法によって良い特徴量が得られ，最適な汎化性能を達成できる
    - ICLR2021最優秀論文賞（投稿論文2997編中トップ8編，0.3%）
  - トランスフォーマーより計算効率の良い代替モデルの探索
  - 事後学習の新手法の提案とテスト時推論の効率性解明：
    - 拡散モデルを事後学習手法，大規模言語モデルのアラインメント手法
    - 文脈内学習の学習原理を解明
    - 思考連鎖による指数関数的な学習効率の向上
- 国内外での評価：
  - 基調講演: ACML2022, ALT2023, AISTATS2026
  - チュートリアル: ACML2021, MLSS2024, ISIT2024, ICONIP2025, CPAL2026
  - 文部科学大臣表彰，日本学術振興会賞，東京大学総長大賞，
    日本神経回路学会論文賞，シンガポールAIVPグラント（1.6億円/3年）

連続最適化（武田TD）

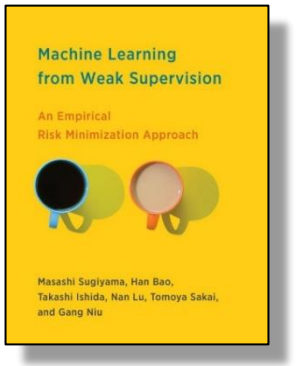
- 求解困難な最適化問題に対して，理論保証付き効率的解法を提案
  - AI for Scienceで使われる非凸スパース最適化法の学習高速化
    - 問題点：スパース正則化関数，深層学習の非平滑活性化関数などの非凸非平滑性が理論的・実際的に効率的な最適化法の構築を阻む
    - 2018年に近接DC最適化法(pDCA)を提案し，L0スパース正則化問題に対する効率的解法として定着
  - 敵対的攻撃対策やハイパーパラメータ調整に役立つ意思決定多段階化
    - 問題点：最も単純な2段階最適化問題に対して超勾配を用いた効率的解法が提案されたが，他ケースについては依然として求解困難
    - 非平滑2段階問題，リーマン2段階問題，多段階最適化問題へと拡張．いずれも，収束保証のある勾配ベースの効率的解法を初めて提案
    - 連続最適化分野最大国際会議ICCOPT2022（3年おき開催）のsemi-plenary講演者として本成果報告
  - ランダム行列理論に基づく巨大学習モデルの極小化
    - 問題点：特別な構造をもたない高次元最適化問題の解法研究は長らく大きな進展がない．
    - 2019年頃から，通常の非線形最適化法にランダム射影技法を組み込む解法構築が始まり，8本の論文を出版
    - 本成果をうけて，数理最適化分野における最大の国際会議ISMP2027（3年おき開催）の「Random Methods for Continuous Optimization」のStream Organizerに選出

以下，主要な研究成果の詳細です

不完全情報学習チーム

- 既存の機械学習の弱点3つを克服する革新的な新技術を開発

1. 大量の教師データが必要：
   - 弱い教師情報からでも学習できる新技術を開発
   - MIT Pressより全貌をまとめた英語の専門書を出版
   - 主要論文の累計被引用数は4000+回



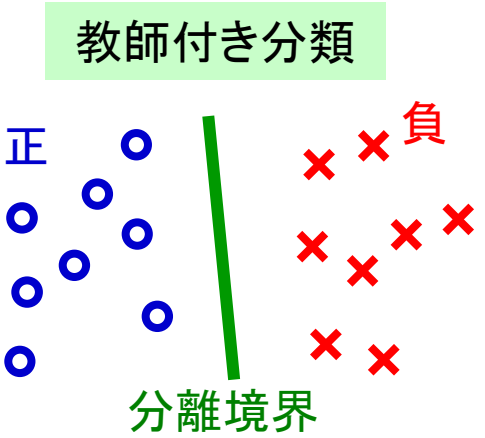
2. 正確な教師データが必要：
   - 雑音を含む教師情報からでも学習できる新技術を開発
   - NeurIPS'18発表論文は被引用3000+回（主要論文累計6800+回）

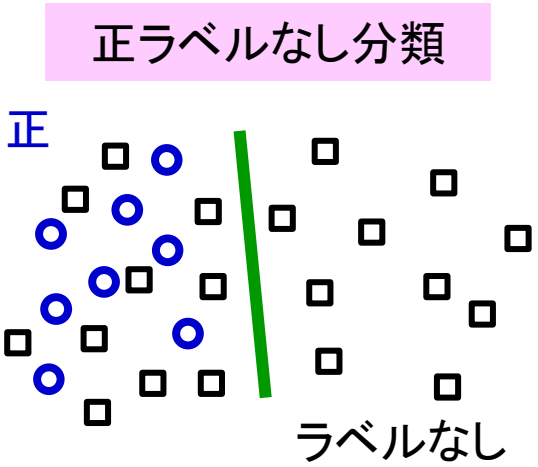
3. バイアスのない学習データが必要：
   - 15年前に開発した転移学習技術が生成AIの基盤技術としてリバイバル
   - 任意のバイアスに対応できる革新技術へと発展
   - 主要論文の累計被引用数は12100+回

- 機械学習3大国際会議の一つ
  ICLR2023にて日本人初の基調講演

成果1：弱教師付き学習

- 教師付き分類：
  - 大量の良質な教師データを用いることにより，人間と同等かそれ以上の予測性能を達成：
  - 画像理解，音声認識，機械翻訳・・・
- しかし，応用分野によっては，教師データを簡単に取れない：
  - 医療，自然災害，材料，プライバシ・・・
- 容易に集められる「弱い」教師情報を活用したい！
  - 例：正例とラベルなしデータからの分類
- 提案した2つの技法：
  - 分類誤差を弱い教師情報から不偏推定
  - 不偏推定量の系統的な非負補正



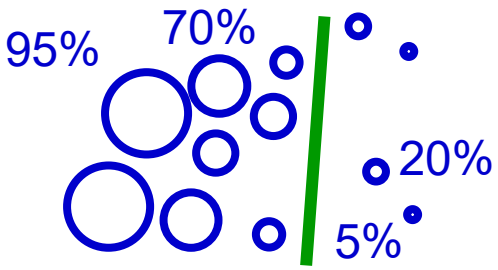


du Plessis+ (NeurIPS2014, ICML2015, MLJ2017), Niu+ (NeurIPS2016), Kiryo+ (NeurIPS2017)

例：クリック予測

様々な弱教師付き分類への拡張

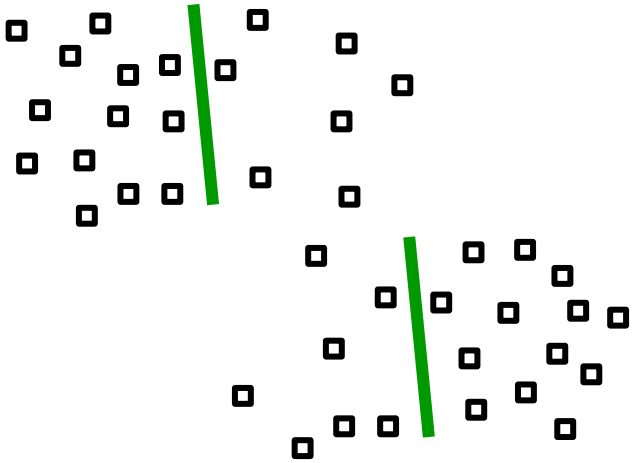
正信頼度学習



Ishida+ (NeurIPS2018), Shinoda+ (IJCAI2021)

例：購買予測

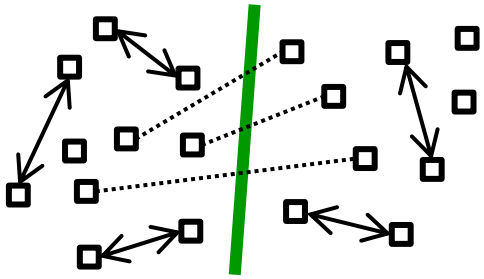
ラベルなしラベルなし分類



du Plessis+ (TAAI2013), Lu+ (ICLR2019, AISTATS2020), Charoenphakdee+ (ICML2019), Lei+ (ICML2021)

例：異なる母集団からの学習

類似非類似ラベルなし分類



Bao+ (ICML2018), Shimada+ (NeCo2021), Dan+ (ECMLPKDD2021), Cao+ (ICML2021), Feng+ (ICML2021)

例：機微情報予測

- 多クラス分類へも拡張可能：
  - 誤ったラベル，曖昧なラベル・・・

Ishida+ (NeurIPS2017, ICML2019), Chou+ (ICML2020), Feng+ (ICML2020, NeurIPS2020), Lv+ (ICML2020), Cao+ (arXiv2021)

- 任意の損失，分類器，最適化法，正則化に適用可能！
- さらなる発展：
  - 統一的枠組み，新しい問題設定，新しい手法・・・

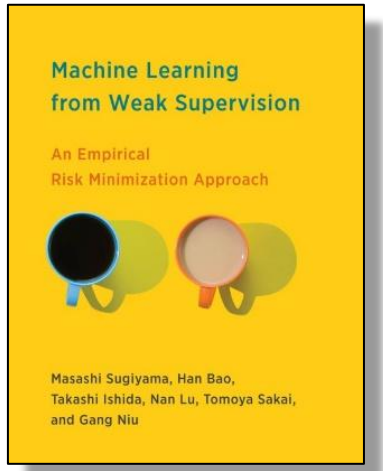
Chiang+ (TMLR2025), Chen+ (ICML2024), Lv+ (NeurIPS2024) Wang+ (NeurIPS2023,ICML2024, ICLR2025),

  - 音響信号処理，強化学習・模倣学習，大規模事前学習モデル・・・

Ito & Sugiyama (ICASSP2023, Best Paper Award), Cai+ (NeurIPS2023), Nishimori+ (RLC2025), Zhang+ (ICML2024), Li+ (MLJ2025)

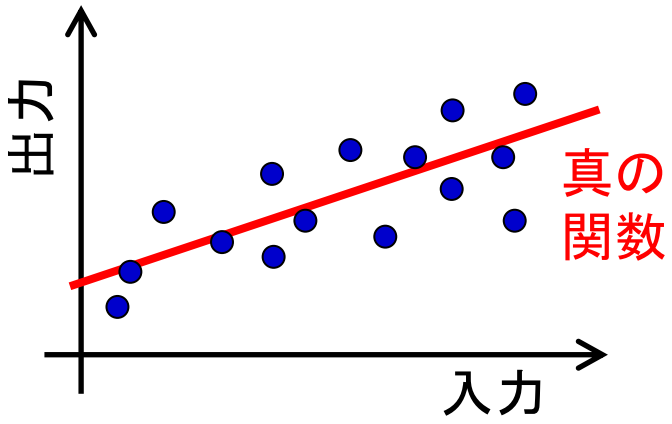
「経験リスク最小化に基づく弱教師付き学習」

Sugiyama, Bao, Ishida, Lu, Sakai & Niu, Machine Learning from Weak Supervision, MIT Press, 320 pages 2022.

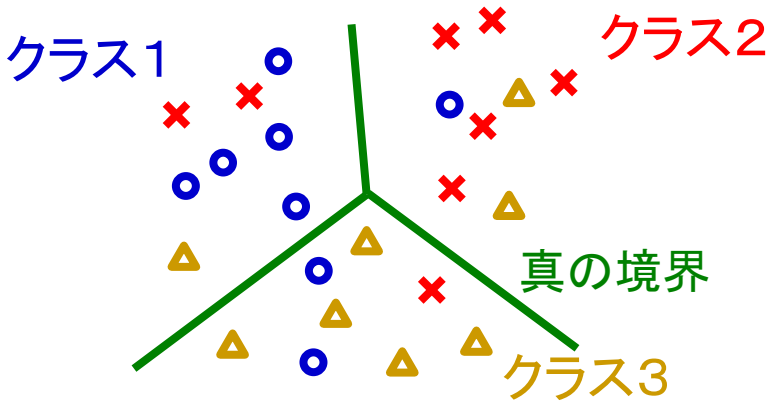


成果2：教師雑音ロバスト学習

回帰（加法雑音）



分類（ラベル反転）



- **回帰**：単にデータを増やせばOK（一致性がある）
- **分類**：データを増やしてもダメ（一致性がない）
  - 明示的な**雑音除去機構**が必要！

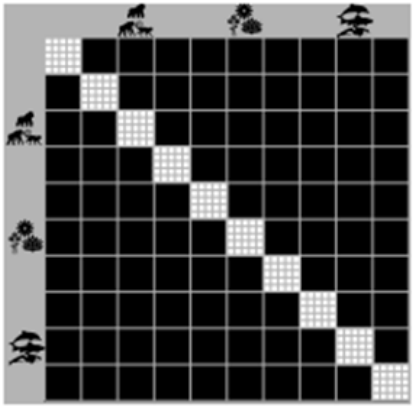
ラベル雑音の補正

- **雑音遷移行列**:
  - ラベル y が $\bar{y}$ に反転する確率を表す行列
  - これが分かれば雑音の影響を補正できる

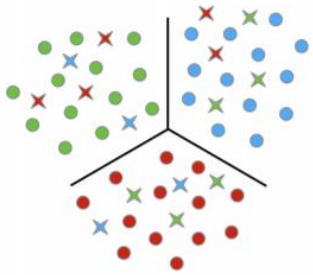
Patrini+ (CVPR2017)

y \ $\bar{y}$			
	1	0	0
	0.1	0.8	0.1
	0.5	0.5	0

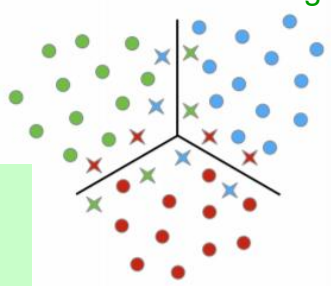
- **雑音を含むデータから雑音遷移行列を推定**:
  - ヒトの認知バイアスを活用 — Han+ (NeurIPS2018)
  - 推定誤差の低減 — Xia+ (NeurIPS2019), Yao+ (NeurIPS2020)
  - 分類器との同時推定 — Zhang+ (ICML2021)
  - 弱い仮定のもとでの一致推定 — Li+ (ICML2021)



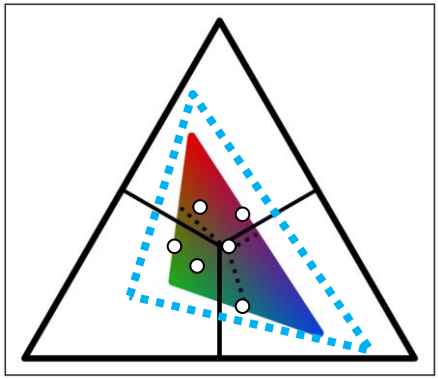
- **入力依存雑音**への拡張 — Xia+ (NeurIPS2020), Berthon+ (ICML2021), Cheng+ (CVPR2022)



入力非依存



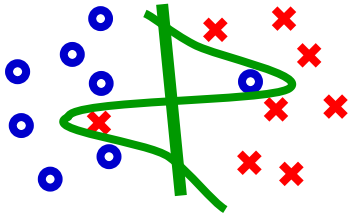
入力依存



共教示

ニューラルネットの記憶能力： Arpit+ (ICML2017), Zhang+ (ICLR2017)

- 確率的降下学習は雑音なしデータを早く記憶
- しかし，単純な早期終了ではうまくいかない



2つのニューラルネットを用いた共教示：

- 誤差の小さいデータを選んで教え合う

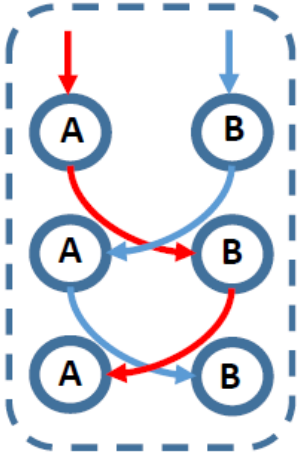
Han+ (NeurIPS2018)

- 出力が合致しないデータだけを教える

Yu+ (ICML2019)

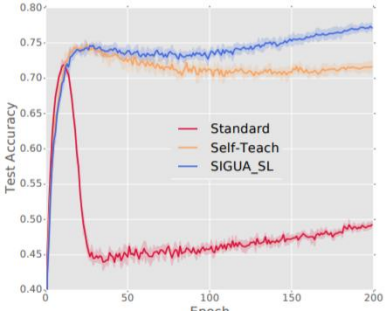
- 誤差の大きいデータに対して勾配上昇

Han+ (ICML2020)



理論はないが，実験的には超ロバスト：

- 50%のラベルをランダムに変えても大丈夫！

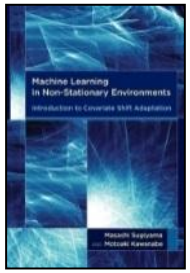


成果3：転移学習

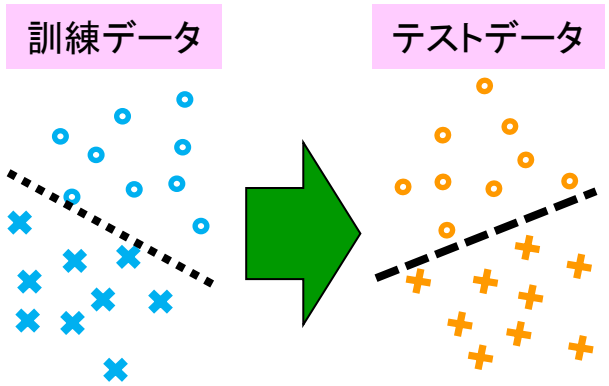
- 訓練データとテストデータの分布が異なると，標準的な機械学習法はうまくいかない：
  - 環境の時間変化
  - 標本選択バイアス（プライバシ）
- 転移学習：訓練データをテスト環境に適応（転移）させる
- 典型的な設定：共変量シフト
  - 入力分布だけが変化 Shimodaira (JSPI2000)
- 基本技術：重要度重み付き学習

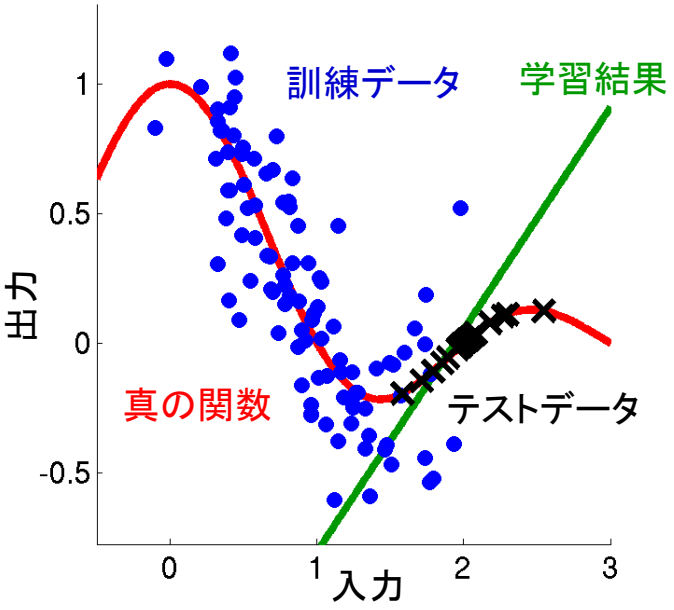
$$\underset{f\in\mathcal{F}}{\operatorname{argmin}}\left[\sum_{i=1}^{n_{\mathrm{tr}}}\frac{p_{\mathrm{te}}(\boldsymbol{x}_i^{\mathrm{tr}})}{p_{\mathrm{tr}}(\boldsymbol{x}_i^{\mathrm{tr}})}\ell(f(\boldsymbol{x}_i^{\mathrm{tr}}),y_i^{\mathrm{tr}})\right]$$

$\boldsymbol{x}$：入力　　y：出力



Sugiyama & Kawanabe
(MIT Press 2012)

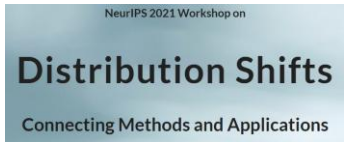


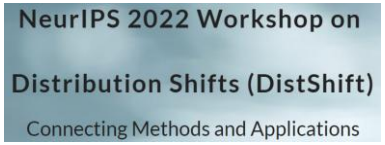


最近の発展

近年再注目：

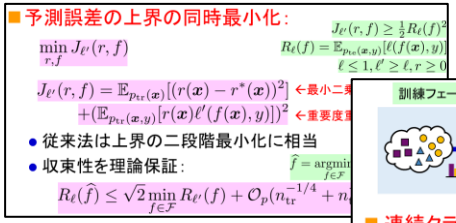
- NeurIPS2021-2025 ワークショップ他

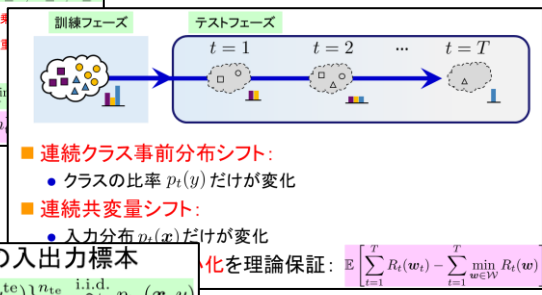


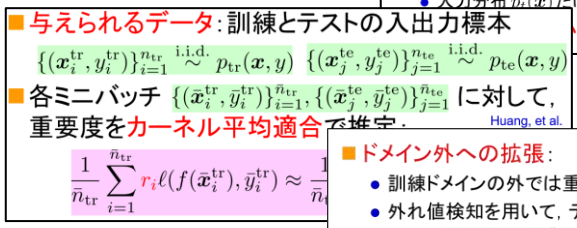


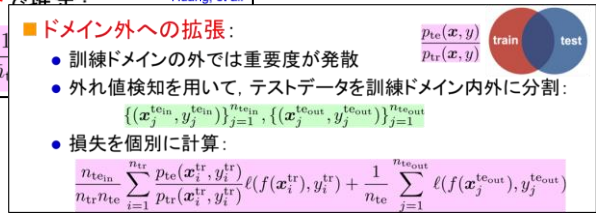
新しい技術：

- 重みと予測器の同時学習
  Zhang+ (ACML2020 Best Paper Award, SNCS2021)
- 連続分布シフトへの拡張
  Bai+ (NeurIPS2022), Zhang+ (NeurIPS2023, ICML2025), Qian+ (ICML2024)
- 同時分布シフトへの拡張
  Fang+ (NeurIPS2020 Spotlight)
- 分布外適応への拡張
  Fang+ (NeurIPS2023 Spotlight)









大規模言語モデルへの応用：

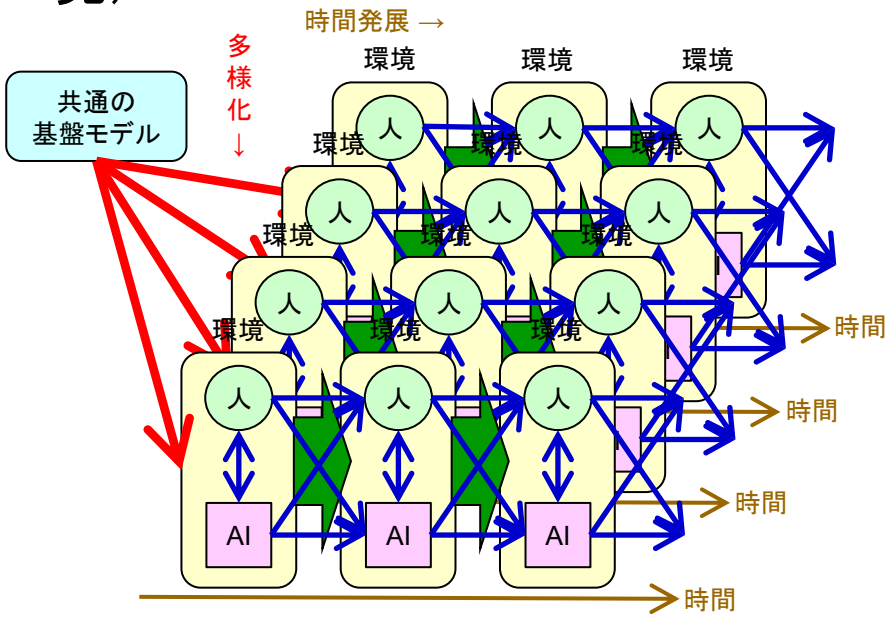
- アラインメント（人間の好みに合わせる） Ackermann+ (COLM2025)
- 運用時シフト適応 Lodkaew+(TMLR2025)

今後の展望

- 基盤モデルの個別化に伴う機械学習：
  - 時間軸に沿った学習
  - 時間軸と多様軸を組み合わせた学習

【理研内連携の例】
・Zhao・Minh：PRI・柚木
・鈴木・園田：iTHEMS・石川、坂上
・清水：CSRS・白須、市橋
・荒井：iTHEMS・清田、R-IH・吉野
・Khan：BDR・泰地
・坂内：iTHEMS・田中
・Huang：R-CCS・Wahib、RQC・Nori

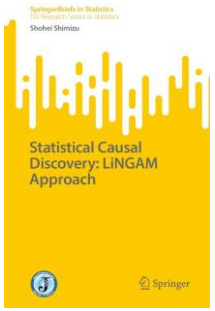
- 具体的なテーマ：
  - 不完全情報学習，構造学習（杉山，河原）
  - オンライン学習，逐次意思決定（伊藤，畑埜）
  - 不確定性定量化，近似推論（Khan，二見）
  - 事後推論（鈴木，今泉）
  - 因果推論（清水，星野）
  - 多エージェント学習（坂田，五十嵐）
  - 大規模データ・大規模モデル学習（武田，Zhao，田部井）
  - 数理基礎（坂内）



因果推論チーム

- AIをさらに発展させる因果推論の新技術を開発
  - AI for Scienceが盛り上がる今こそ、因果推論は不可欠

1. 従来 因果構造が既知である必要:
   - 潜在共通原因があってもデータ駆動で因果構造を探索する新技術を開発
   - Springerより一連の研究をまとめた英語の専門書を出版



2. 従来 科学者・実務家の専門知識が必要:
   - LLMと統計的因果構造探索の連携
   - 全国規模の保険者データベースによるPoCの実施
     - JST CREST (信頼されるAI領域)

3. 従来 科学者・実務家用ソフトウェアの不足:
   - Pythonパッケージの開発2つ: 月2万・10万ダウンロード以上
   - 商用ソフトウェア採用3件 (1件監修)

Causalas

- 因果推論専門国際WS・会議での招待講演・パネリスト
  - ワークショップ at UAI2023, KDD2021, NeuIPS2020
  - 会議 Pacific Causal Inference Conference 2024, 2020

因果探索

- 因果グラフを描く「支援」
- データを用いて因果グラフを推測するための方法論
- 領域知識以外の手段

仮定(+領域知識)

- 未観測共通原因の有無
- 非巡回 or 巡回
- 分布
- 関数形など

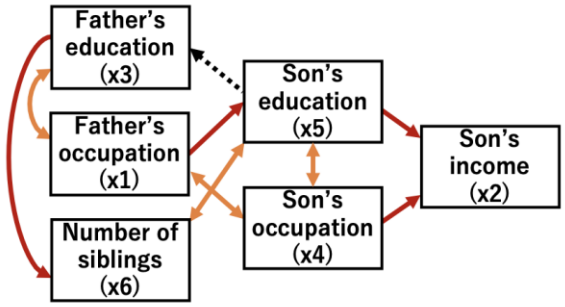
＋

データ

x1	x2	x3	x4	x5	x6
87.9	45433	17	76.3	17	1
87.9	55071	16	86	18	2
62.1	113159	16	87.9	16	0
78.5	30289	16	30.1	14	4
32.3	113159	20	63.5	20	7
60.6	55071	17	83.7	17	1
76.4	55071	16	78	14	2
63.5	37173	12	63.2	16	3
63.2	113159	14	86.5	17	1
36.5	37173	12	83.7	12	4

推測

因果グラフ

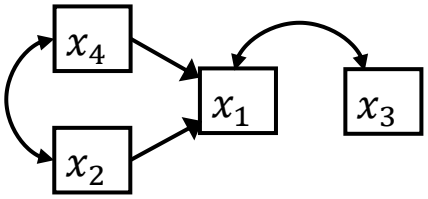


Maeda and Shimizu (2020)

1. 潜在変数のある場合の因果探索

- 非線形で未観測共通原因がある場合
  - Maeda and Shimizu (UAI2021, 2024): モデル・推定法の提案
  - Pham, Maeda, Shimizu (AISTATS2026): 十分条件とSound and complete algorithm

$$x_i = \sum_{j:\, x_i\text{の親}} f_{ij}(x_j) + \sum_{k:\, x_i\text{の親}} g_{ik}(u_k) + e_i$$



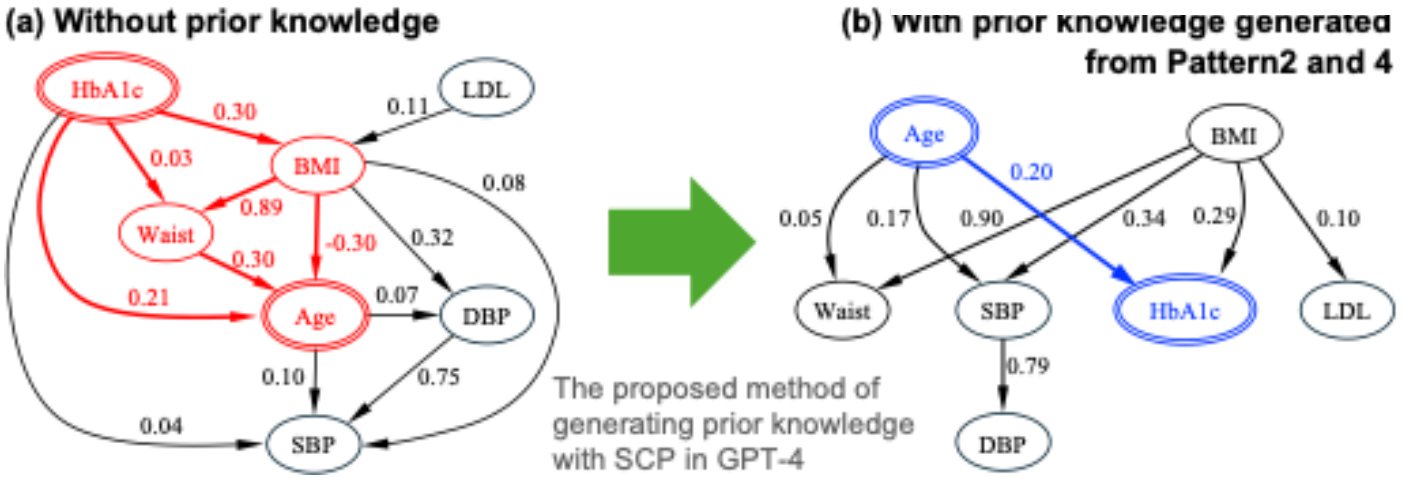
向き・交絡の「存在」が識別可能

2. LLM × 因果探索

- Causal parrots (Zečević+2023): オウム返し
- LLMで背景知識を収集・データから因果探索
  (Takayama+2025)

背景知識なし

LLMによる背景知識あり



データが少ないと失敗

リークのない (LLMが知らない) 健康診断データで評価

3. ソフトウェア

- lingamとcausal-learnをgithubにて公開
  - 開発した手法や機能を実装する因果探索パッケージ
  - PyPI Statsによる過去1ヶ月DL数(2/4現在)
    - lingam: 22,258
    - causa-learn: 104,442 (PyWhyのレポジトリ)
- ノーコードツール
  - Causalas (SCREENアドバンストシステムソリューションズ)
  - Node AI (NTTドコモビジネス)
  - NTech Predict (neutral)

  - Causal analysis (hootfolio)
  - CALC (ソニー)

将来展望:「自律」因果探索へ

- 従来の課題
  - 背景知識収集やそれに基づく変数定義、どのデータを使うかの判断は“手作業”
- 自律因果探索
  - AIが知識・データを自動統合し、因果構造＋変数定義の因果仮説を探索
  - シミュレーション・自動実験による検証で効率化とスケール拡大
- 未来像：因果がわかるAIサイエンティスト
  - AIと人間の協働
  - 医学・社会科学、創薬科学・材料科学など
    広く科学の発見や理解・産業応用を加速
  - AI for Science: 自律型でも支援型でも
  - Science for AI: 因果推論できるAIへ

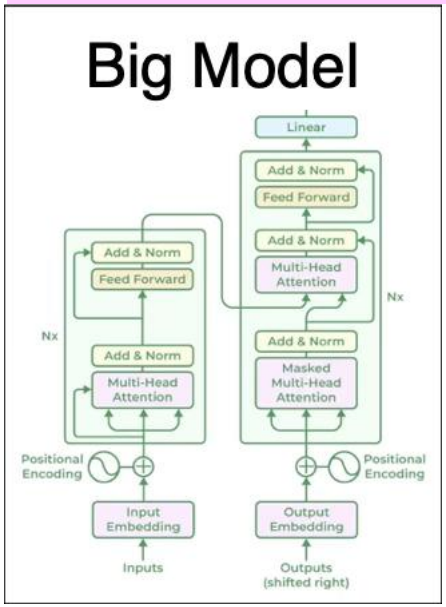
Adaptive Bayesian Intelligence Team

- Sustainable AI: drastically reduce the cost of AI training
  - Break free from dependence on huge data and computing centers
  - Instead, continually learn and quickly adapt (just like children)
- Main Achievement: The Bayesian Learning Rule
  - It unifies many popular algorithms and can help us design better ones
  - One of our new algorithm won a NeurIPS-2021 challenge
  - It also led to the first Bayesian algorithm to successfully train LLMs
  - Using this, we are now working on continual learning to reduce cost
- Awarded JST-CREST grant of 220M JPY on this topic
- Invited Speaker at leading venues in AI and Bayes
  - NeurIPS 2019 tutorial (one of the most popular, #audience >7K)
  - Keynote at ISBA 2026 and BayesComp 2025 (two largest Bayesian conferences), EurIPS 2025 (#audience >2k), CoLLAs 2024
- Served as Program/General Chairs in leading ML venues
  - ICLR 2024 (largest deep learning conference), AISTATS 2025, 2026

Unsustainable AI Training

- AI training requires a huge amount of data and compute, as well as energy (electricity, etc.)

Reason: stochastic training need access to the whole model and data at all times



Stochastic training

$$\theta \leftarrow \theta - \rho \sum_{j \in \mathcal{B}} \nabla \ell_j$$

Parameter

Loss gradient

Minibatch

- Our goal: break free from this dependence.
- Human–like adaptation: Learn continually by selecting relevant data and parameters

Bayesian Learning Rule (BLR) [1,2]

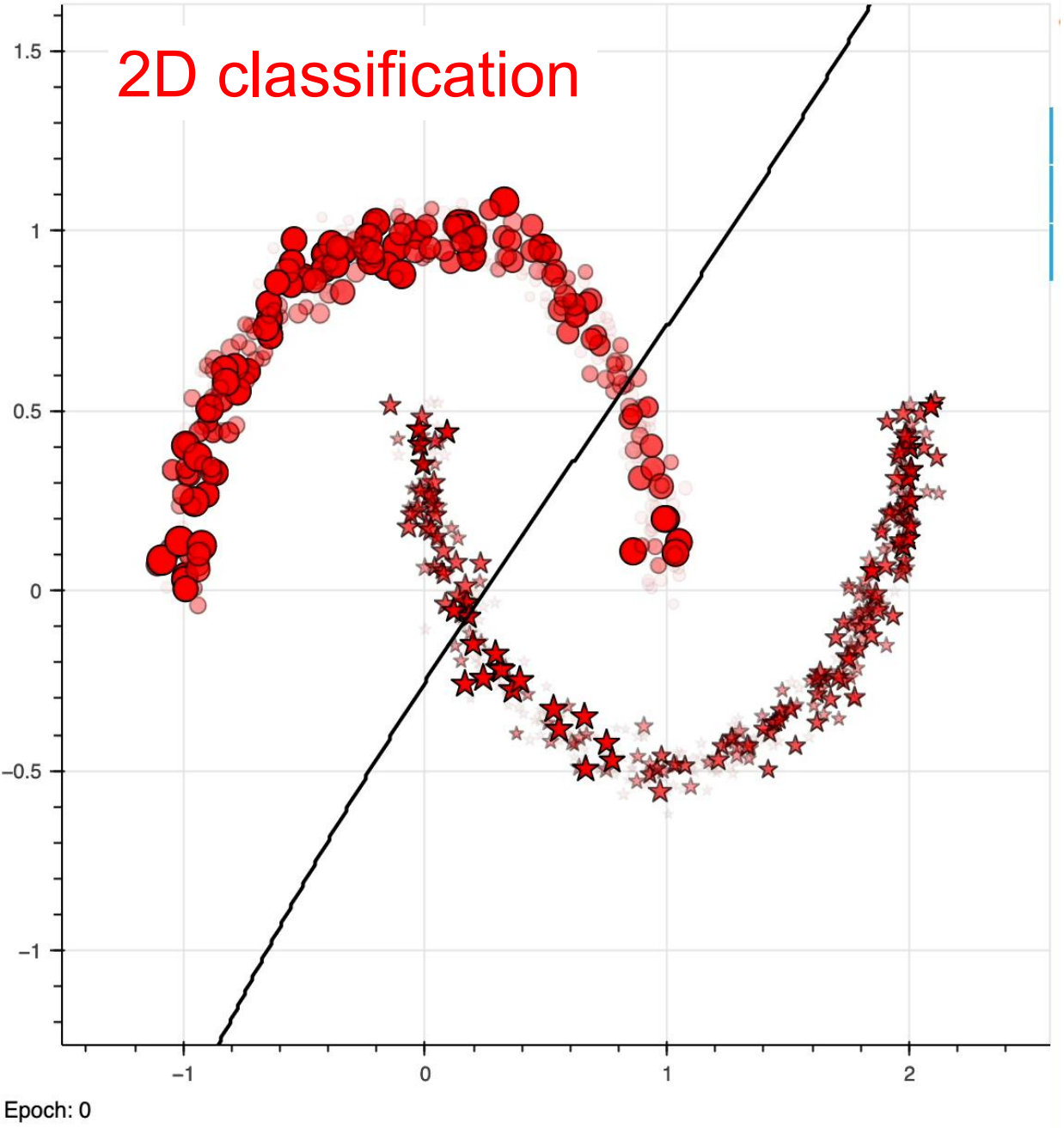
- We use "uncertainty" to decide where to focus
- Uncertainty is obtained with the Bayesian learning rule which unifies many algorithms

All these algorithms are special case of the Bayesian Learning Rule

Optimization	Deep-Learning	Approximate Inference
Gradient Descent	SGD, RMSprop and Adam	Conjugate Bayes
Newton's Method	Sharpness-Aware Minimization	Laplace's Method
Multimodal Optimization	Dropout, STE, Label Smoothing	Expectation Maximization
	Shampoo….	Stochastic Variational Inference
		Variational Message Passing

- The BLR helps us understand and "focus" on the sources of uncertainties to adapt quickly

1. Khan and Rue, The Bayesian Learning Rule, JMLR (2023)
2. Khan and Lin. Conjugate-Compute Variational Inference, AISTATS (2017)



More relevant examples are shown with bigger markers.

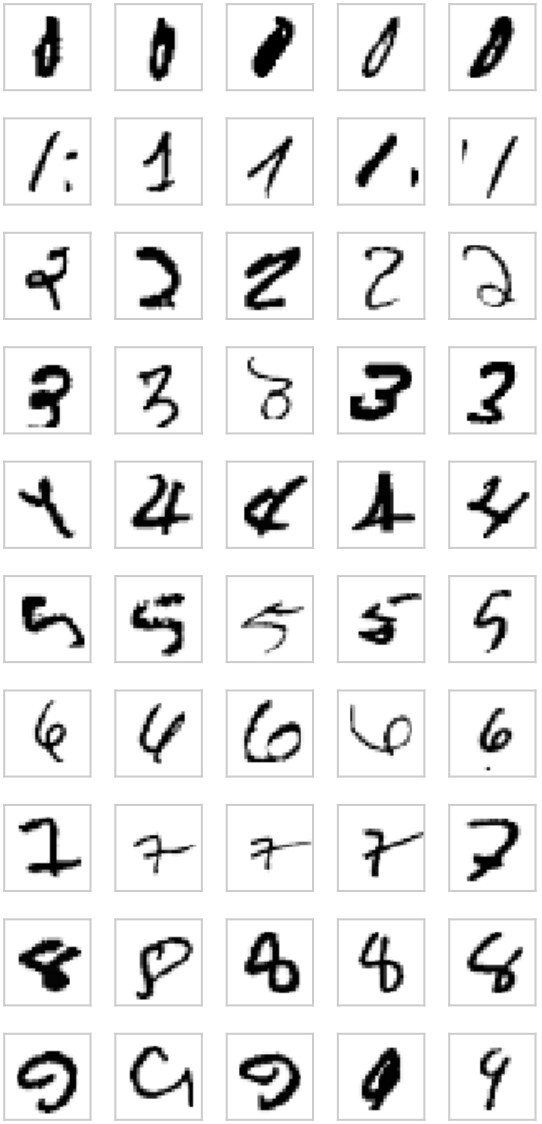
Cost can be reduced by focusing on such relevant examples and ignoring others.

Trained with the BLR variant called IVON.

1. Nickl, Xu, Tailor, Moellenhoff, Khan, The memory-perturbation equation, NeurIPS (2023)

An Example for MNIST

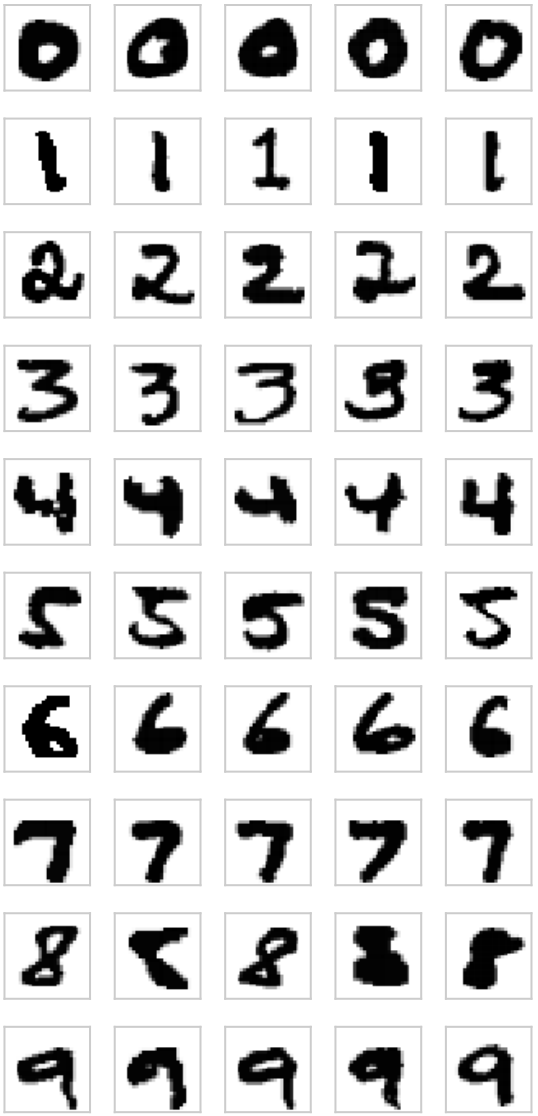
More Relevant



Training should focus on difficult cases (left), not that much on the easy cases (right).

We can figure this out for many existing algorithms by using our Bayesian framework.

Less Relevant



Learning to Focus

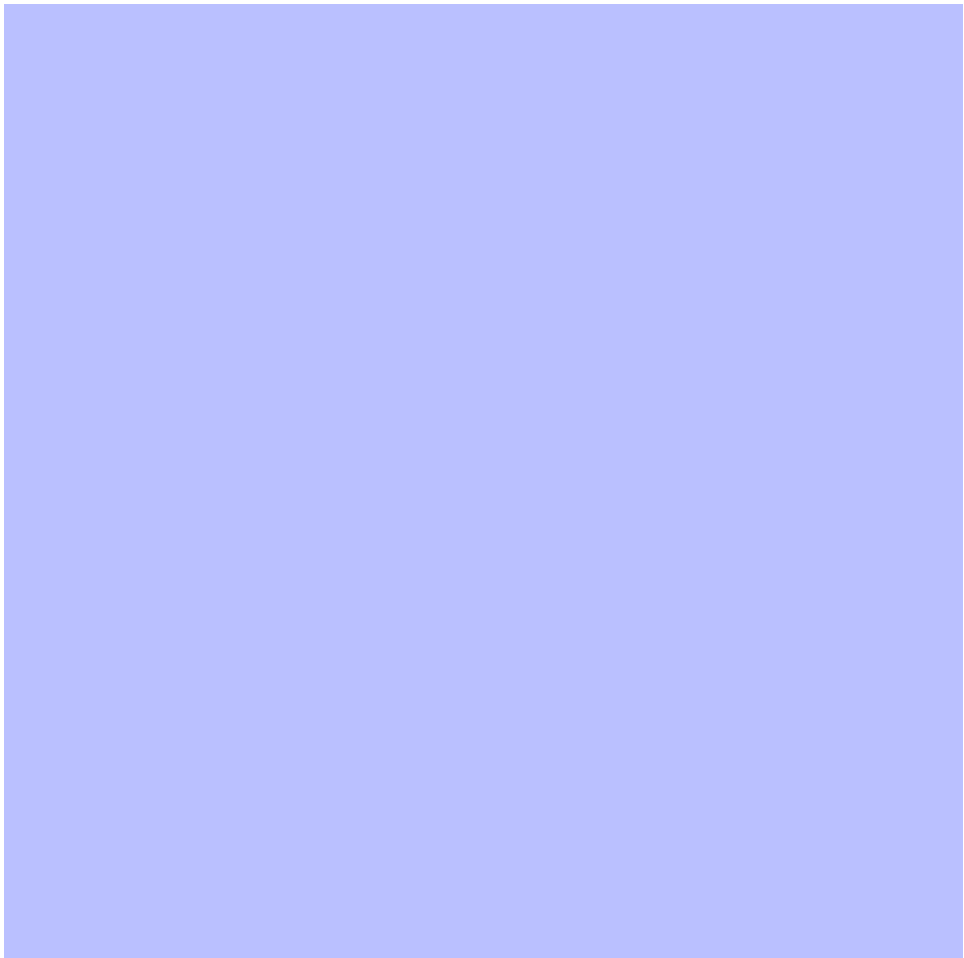
We are working on a new PoCo optimizer that learns by focusing on a few examples (shown in black), as opposed to Adam that need access to all examples all the times

The Adam optimizer

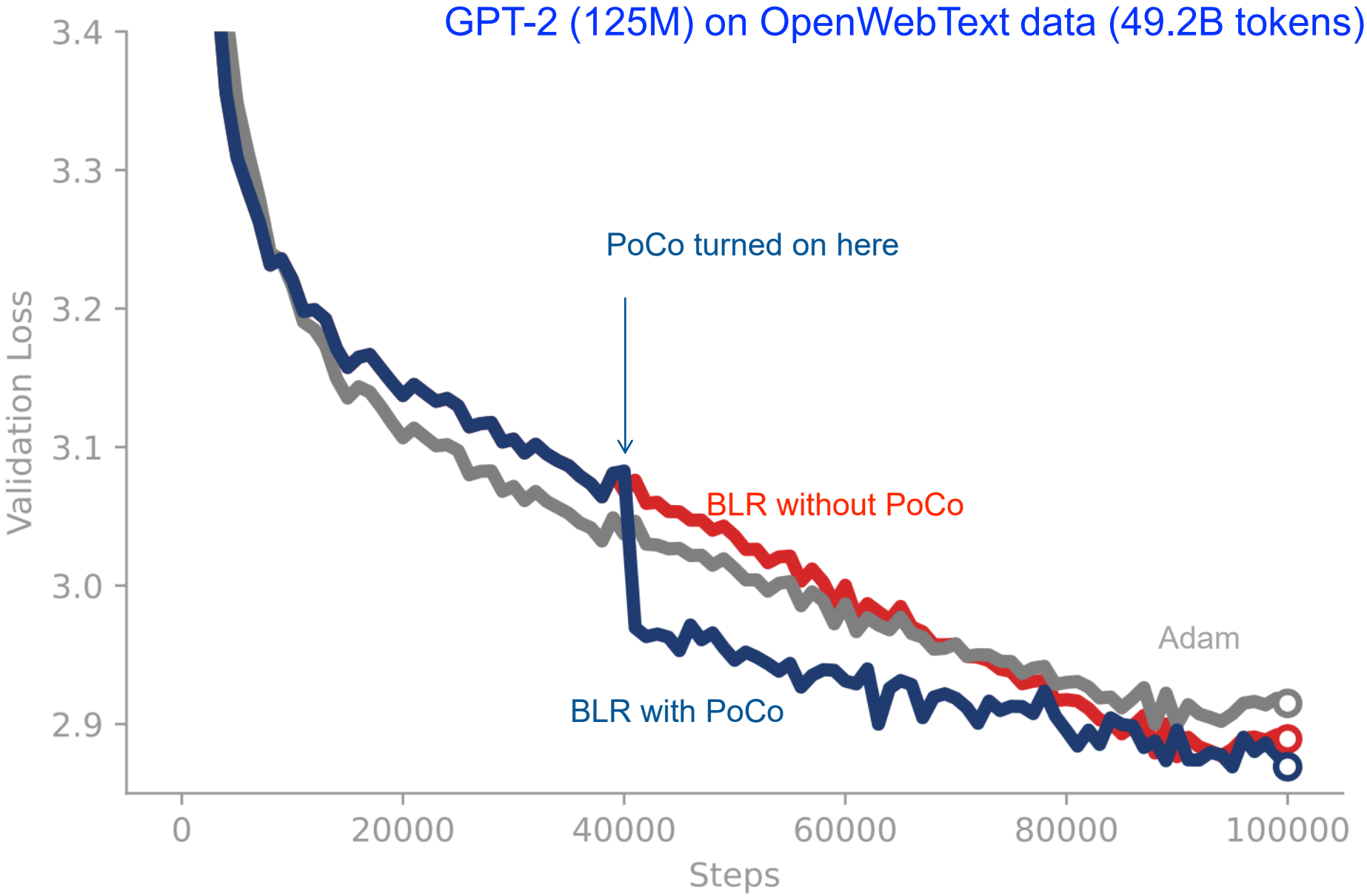
Recently picked examples are in large markers

Our new Posterior Correction (PoCo) optimizer

Relevant examples shown in black



PoCo can boost LLM performance



成果まとめ（深層学習理論チーム）

- 深層基盤モデルの表現力の解明
  - **DNNは次元の呪いを回避する**
    - 非等方Besov空間; γ-平滑関数空間
  - **深層基盤モデルの最適性を証明**
    - 拡散モデルは全ての学習法の中でほぼ最適な推定精度を達成, データの低次元構造を自動的に捉えて次元の呪いを回避
    - Transformerが長い文章を扱える理由⇒重要なトークンを優先的に選択する
- ニューラルネットワークの最適化原理の解明
  - **勾配法による「良い」特徴量が得られることを証明**
    - 汎化性能の向上(最適性)を理論的に解明
    - 非特徴学習法よりも学習時間を短くできることを解明
  - **Neural Tangent Kernelを用いた解析でSGDの最適性を解明 (ICLR2021 outstanding paper award)**
- トランスフォーマの代替モデルの探索
  - **Attentionの計算量を改善する手法**としてSSMの表現力を解析
    - 多層にすればAttentionと同様の表現力を持つことを解明
  - 非線形特徴量による**線形Attentionの次元選択手法**の開発
    - 層ごとに最適な特徴次元を効率的に計算可能
- 事後学習の新手法の提案とテスト時推論の効率性解明
  - **拡散モデルを正しく事後学習する方法 /人間の選好モデルを使わないLLMのアラインメント手法**
  - **文脈内学習**の学習原理を解明 (非線形関係の学習，ベイズ推論との関連を解明)
  - **思考連鎖による指数関数的な学習効率の向上**を証明
- 招待講演: (1)基調講演: ACML2022, ALT2023, AISTATS2026, (2)チュートリアル: ACML2021, MLSS2024, ISIT2024, ICONIP2025, CPAL2026, (3) 国際ワークショップ: Simons Institute, Oberwolbach, CIRM, EPFL,...
- 文部科学大臣表彰，日本学術振興会賞，東京大学総長大賞，日本神経回路学会論文賞，AIVP グラント 1.6億円/3年

深層学習の最適性

深層学習は特徴学習により学習効率を向上: ミニマックス最適性
Besov空間 & 混合平滑Besov (Suzuki, ICLR2019)

スパース性と凸包の理論：統一理論
- Hayakawa & Suzuki (2020), 日本神経回路学会論文賞

深層学習が優越する理由を凸包の幾何学で統一的に説明

特徴学習で次元の呪いを回避
- Anisotropic Besov space (Suzuki & Nitanda, NeurIPS2021)
- γ-smooth function class (Okumoto & Suzuki, ICLR2022)
- Infinite dimensional input-output (Nishimura & Suzuki, ICLR2023)

滑らかでない場所を特定して狙い撃ち
- Variable smooth Besov (Tsuji & Suzuki, 2021)

無駄な情報を削り情報を圧縮することで汎化性能向上

近年の生成基盤モデルの最適性

拡散モデル
(Oko,Akiyama&Suzuki, ICML2023)

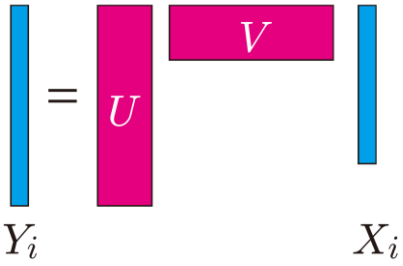
Transformer
(Takakura&Suzuki, ICML2023)

なぜ深層学習が良いのか？

統計的決定理論におけるミニマックス最適性理論で特徴づけ可能

縮小ランク回帰

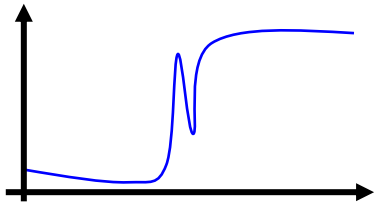
特徴空間の次元が低い状況は深層学習が得意



Besov空間

[Suzuki, 2019]

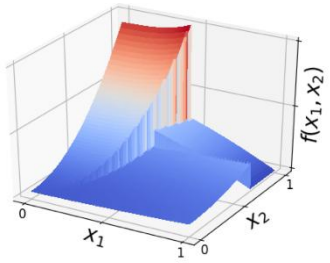
滑らかさが非一様な関数の推定は深層学習が得意



区分滑らかな関数

[Imaizumi&Fukumizu, 2019]

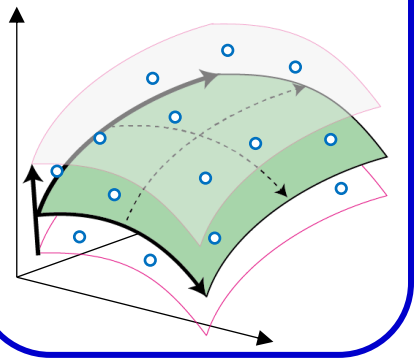
不連続な関数の推定は深層学習が得意



低次元データ

[Schmidt-Hieber, 2019] [Nakada&Imaizumi, 2019][Chen et al., 2019][Suzuki&Nitanda, 2019]

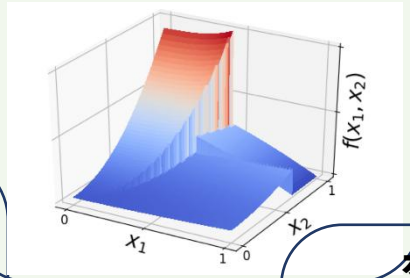
データが低次元部分空間上に分布していたら深層学習が有利



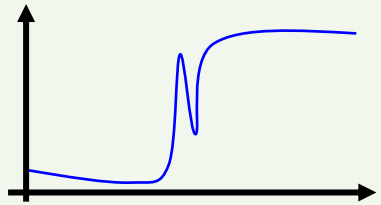
推定精度				
深層	$\frac{r(M+N)}{n}$	$n^{-\frac{2s}{2s+d}}$	$n^{-\frac{2s}{2s+d}} \vee n^{-\frac{\alpha}{\alpha+D-1}}$	$n^{-\frac{2s}{2s+D}}$
カーネル	$\frac{MN}{n}$	$n^{-\frac{2s-2d(1/p-1/2)_+}{2s+d-2d(1/p-1/2)_+}}$	$\frac{1}{\sqrt{n}}$	$n^{-\frac{2(s-D/p+d/2)}{2(s-D/p+d/2)+d}} \vee n^{-\frac{2s}{2s+D}}$

数学的一般化

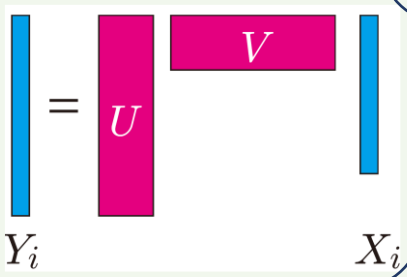
区分滑らかな関数



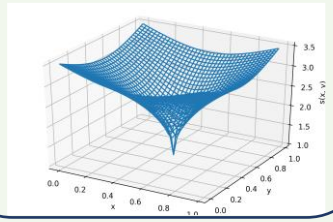
Besov空間



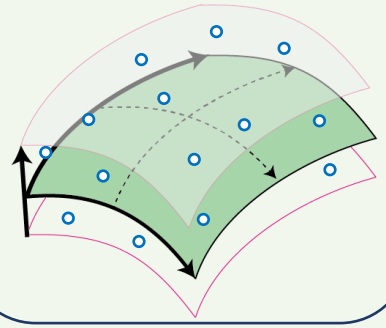
縮小ランク回帰



変動指数
Besov空間



低次元データ



非凸性
スパース性

[Satoshi Hayakawa and Taiji Suzuki: On the minimax optimality and superiority of deep neural network learning over sparse parameter spaces. Neural Networks, 2020.]

拡散モデルの統計理論

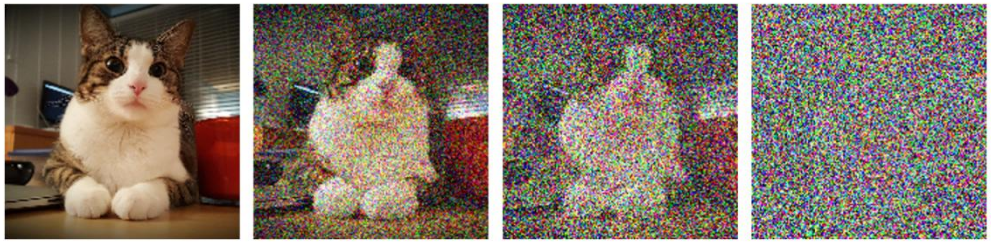
[Kazusato Oko, Shunta Akiyama, Taiji Suzuki: Diffusion Models are Minimax Optimal Distribution Estimators. ICML2023, **oral**]

(2% of all submissions)

Stable diffusion, 2022.

$$\mathrm{d}X_t = -X_t \mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

Forward process



Backward process

$$\mathrm{d}Y_t = (Y_t + 2\nabla \log(p_{\overline{T}-t}(Y_t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad (Y_t \sim X_{\overline{T}-t})$$

経験スコアマッチング推定量:

$$\hat{s} = \underset{s \in \mathrm{DNN}}{\arg\min} \frac{1}{n} \sum_{i=1}^{n} \int_{t=\underline{T}}^{\overline{T}} \mathbb{E}_{X_t|X_0=x_{0,i}} \left[\| s(X_t, t) - \nabla \log p_t(X_t|x_{0,i}) \|^2 \right] \mathrm{d}t$$

定理

Let $\hat{Y}$ be the r.v. generated by the backward process w.r.t. $\hat{s}$, then

$$\mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}} \log^9(n), \quad (s\text{: 密度関数の滑らかさ})$$

$$\mathbb{E}_{D_n}\left[\mathrm{W}_1(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}} \quad (\text{for any } \delta > 0).$$

どちらも（ほぼ）**ミニマックス最適** [Yang & Barron, 1999; Niles-Weed & Berthet, 2022].

(Estimator for W_1 distance requires some modification)

Transformerの推定理論

[Shokichi Takakura, Taiji Suzuki: Approximation and Estimation Ability of Transformers for Sequence-to-Sequence Functions with Infinite Dimensional Input. ICML2023]

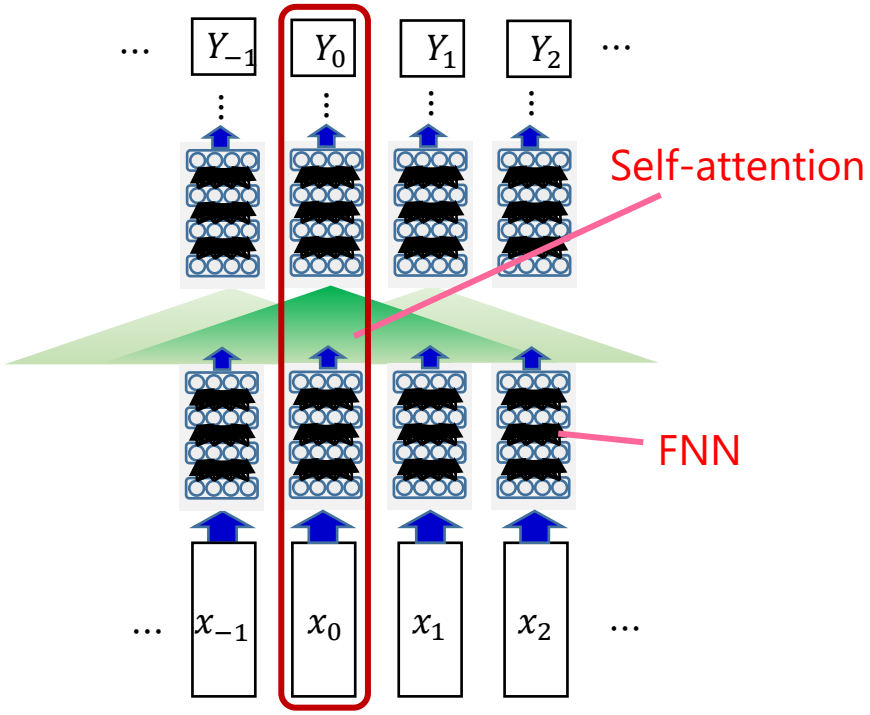
Transformerの性質

- かなり広いトークン幅から重要なトークンを選べる.

→ 次元の呪い？

- 入力に依存して重要なトークンを選択できる.

→ 次元の呪いを回避！



定理 (推定誤差)

$$\frac{1}{r-l+1}\sum_{j=l}^{r}\mathbb{E}[\|\hat{F}_j - F_j^\circ\|_{L_2(P_X)}^2] \lesssim n^{-\frac{2a^\dagger}{2a^\dagger+1}}(\log n)^{2/\alpha+2+\max\{4/\alpha,4\}}$$

(ほぼミニマックス最適)

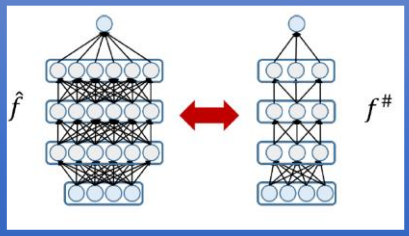
➢ 入力が無限次元でも多項式オーダーの収束レート.

最適化理論と汎化性能

カーネル法による汎化誤差理論: Suzuki (AISTATS2018)

モデル圧縮への応用

- モデル圧縮可能性による汎化性能の解明:
  Suzuki, Abe, Nishimura (ICLR2020)
- Spectral pruning: Suzuki et al. (2020)
- Tensor decomposition for CNN: Jingling et al. (2020)



Neural tangent kernel

- ミニマックス最適性: Nitanda, Suzuki (2021); **ICLR2021 outstanding paper award**

特徴学習と最適化：

- 二重降下現象と初期化の関係: Ba et al. (ICLR2020)
- 前処理付き最適化と汎化誤差への影響: Amari et al. (ICLR2021)
- 勾配法による特徴量の発見と汎化誤差の改善: Ba et al. (NeurIPS2022)

特徴学習と二重降下: Suzuki&Suzuki (ICLR2024), Nishimori&Suzuki (IG2024)

情報指数を用いた詳細な解析

- バッチ再利用によるSGDの最適性: Lee et al. (NeurIPS2024)
- 加法モデルの学習可能性: Oko et al. (COLT2024)
- Mixture of Expertの学習ダイナミクス解明: Kawata et al. (ICML2025)

勾配法とKernel alignment

[Jimmy Ba, Murat A. Erdogdu, Taiji Suzuki, Zhichao Wang, Denny Wu, Greg Yang: High-dimensional Asymptotics of Feature Learning: How One Gradient Step Improves the Representation. NeurIPS2022.]

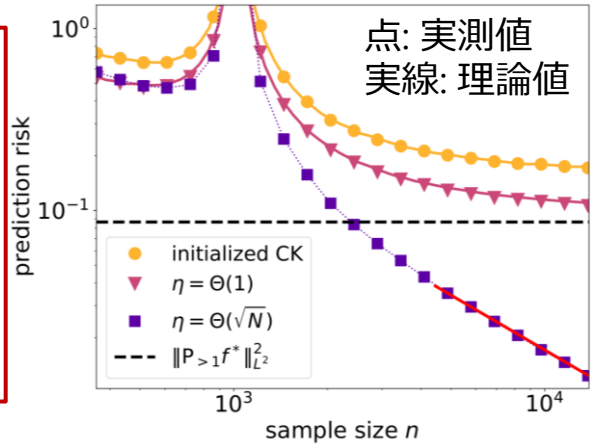
$$f_{\mathrm{NN}}(x) = \frac{1}{\sqrt{N}} \sum_{i=1}^{N} a_i \sigma(\langle x, w_i \rangle) = \frac{1}{\sqrt{N}} a^\top \sigma(W^\top x)$$

問：勾配法でWを更新することで，データに合った特徴量を獲得できるか？
答：大きなステップサイズを用いれば，一回の更新で意味のある特徴量の方向を得ることができる．
→ カーネルAlignment，特徴量学習．

$$W_{k+1} = W_k + \eta \sqrt{N} \nabla L(f_{\mathrm{NN}})$$

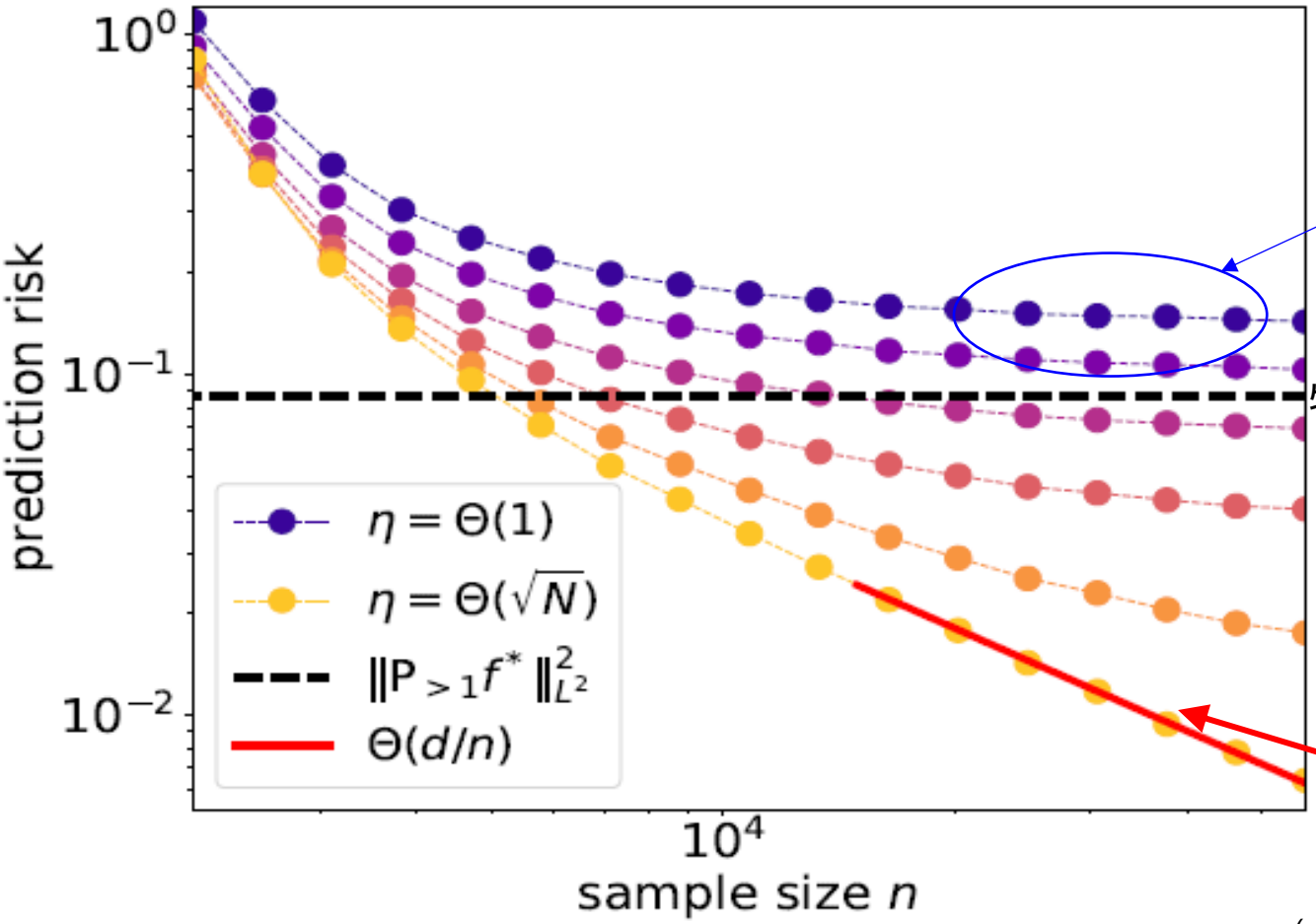
$\boldsymbol{n, d, N \to \infty}$**の極限**を考え，勾配法１回の更新後の予測誤差を評価してみる．

- $\eta = \sqrt{N}$: 大きなステップサイズを用いると，ランダム特徴モデルによるリッジ回帰を優越する．
- $\eta = 1$: 中間的なステップサイズでは横幅無限大のランダム特徴リッジ回帰を優越しないが初期値Wは優越．
- $\eta = o(1)$: 小さなステップサイズでは初期値Wと同じ予測誤差 (NTK-regime)．特徴学習の効果なし．



実験との整合性

(点線：理論値，丸印：実験)



小さなステップサイズ

特徴学習しない場合の限界
（ランダム特徴量）

大きなステップサイズ：
$\Theta(d/n)$で減少

(n:サンプルサイズ，d:入力の次元)

勾配法一回分による更新後の予測誤差．
更新に用いるステップサイズごとにプロット

Single index modelの学習複雑度

- **Gaussian single index model:** $y_i = f_*(x_i) + \varepsilon_i \quad (\varepsilon_i \sim N(0, \varsigma^2))$

$$f_*(x) = \sigma_*(\langle x, \theta \rangle) \qquad (x \sim N(0, I_d))$$

➢ Requires learning the direction $\theta \in \mathbb{R}^d$ and link function σ_*.
- The *informative direction* in only one.
  → **Feature learning**

➢ σ_* is assumed to be a polynomial with degree p and information exponent k.

$$f_*(x) = \sigma_*(\langle x, \theta \rangle) = \sum_{i=k}^{p} \alpha_i^* \mathrm{He}_i(\langle x, \theta \rangle)$$

Information theoretic lower bound (statistical complexity):

Required sample size n:

Kernel method: $n = \Omega(d^p)$ [Ghorbani et al. 19; Donhauser et al. 21; Gavrilopoulos et al. 24;...]

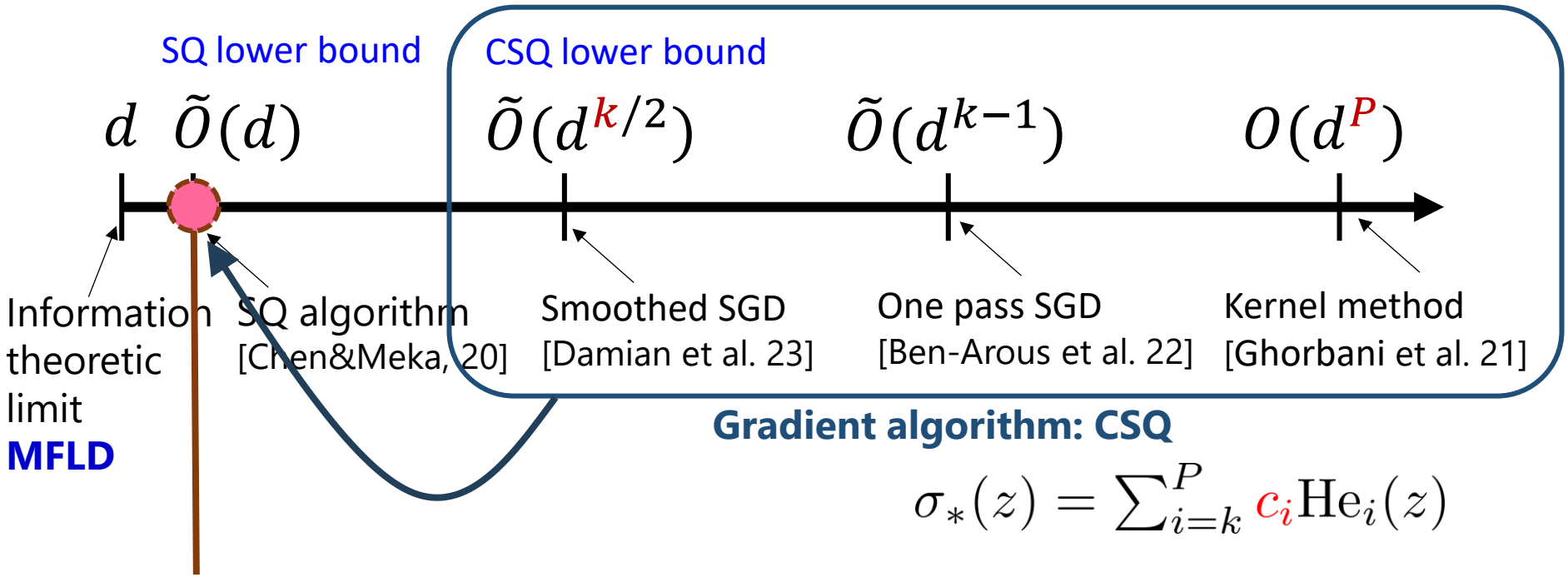
Neural network: $n = \mathrm{O}(d)$ [Bach 17; Barbier et al. 19; Damian et al. 24;...]

- Neural network has better sample complexity due to feature learning ability.
- But, it may require $\exp(d)$ computation.

"Statistical vs Computational" tradeoff?

勾配法によるNNの学習複雑度

$$\nabla_w \mathbb{E}_{x,y}[(y - f_w(x))^2] \propto -\underbrace{\mathbb{E}_{x,y}[y \nabla_w f_w(x)]}_{\text{Correlation statistical query (CSQ)}} + \mathbb{E}_x[f_w(x) \nabla_w f_w(x)]$$



[Dandi et al. 2024]**[Lee, Oko, Suzuki, Wu; NeurIPS2024]**

Gradient method with batch-reuse can implement SQ algorithm so that we obtain $n = \tilde{O}(d)$ sample complexity.

[**Generative exponent**]

特徴学習は計算量も軽減する

Jason D. Lee, Kazusato Oko, Taiji Suzuki, Denny Wu: Neural network learns low-dimensional polynomials with SGD near the information-theoretic limit. NeurIPS2024.

勾配・平均場ランジュバン動力学

平均場ランジュバン動力学の線形収束

- Nitanda, Wu, Suzuki (AISTATS2022)

“二重ループ法”の提案と収束

- PDA: Nitanda, Wu, Suzui (NeurIPS2021)
- P-SDCA: Oko, Suzuki, Wu, Nitanda (ICLR2022)
- Infinite-dim extension: Nishikawa, Suzuki, Nitanda, Wu (NeurIPS2022)

確率的勾配ランジュバン動力学

- SVRG-GLDの解析: Kinoshita, Suzuki (NeurIPS2022)
- 無限次元SGLD: Muzellec et al. (COLT2022)

勾配ランジュバン動力学による最適化と汎化誤差解析 (次元の呪いの回避)

- Infinite-dim GLD: Suzuki (NeurIPS2020), Suzuki, Akiyama (ICLR2021)
- Teacher student: Akiyama, Suzuki (ICML2021,ICLR2023)

Uniform-in-time propagation of chaos（有限粒子近似の導出）:

- Super log-Sobolev inequality: Suzuki, Nitanda, Wu (ICLR2023)
- Convergence analysis with Stochastic gradient/finite particle/discrete time alg. : Suzuki, Nitanda, Wu (NeurIPS2023)

強化学習: Yamamoto et al. (2023); **ミニマックス問題:** Kim et al. (ICLR2024)

Refined PoC: 対数ソボレフ定数に依存しない上界: Nitanda et al. (ICML2025)

確率測度の最適化

$$\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu) = F(\mu) + \beta \begin{cases} \mathrm{KL}(\mu||\mu_{\mathrm{ref}}) \\ \mathrm{Ent}(\mu) \end{cases}$$

応用：ニューラルネットワーク最適化，選好最適化ファインチューニング，ベイズフィルタリング，．．．

- **平均場ランジュバン動力学：** [Nitanda, Wu, Suzuki, AISTATS2022][Suzuki, Nitanda, Wu, NeurIPS2023]

$$\mathrm{d}X_t = -\nabla \frac{\delta F(\mu_t)}{\delta \mu}(X_t)\mathrm{d}t + \sqrt{2\lambda_2}\mathrm{d}B_t$$

線形収束：$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2\alpha\lambda_2 t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*))$

対数ソボレフ不等式

物理 確率

- **拡散モデルのファインチューニング：** [Kawata, Oko, Nitanda, Suzuki, ICLR2025]

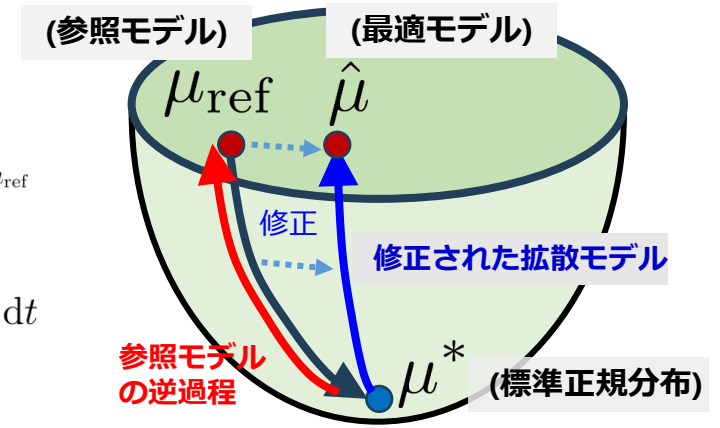
μ_{ref}: 事前学習済み拡散モデル

双対平均化法： For $k = 1, \ldots, K-1$:

$$\mu^{(k+1)} = \arg\min_{\mu\in\mathcal{P}} \frac{2}{k(k+1)} \sum_{j=1}^{k} j \left(\mathbb{E}_\mu \left[\frac{\delta F(\mu^{(j)})}{\delta \mu} \right] + \beta \mathrm{KL}(\mu\|\mu_{\mathrm{ref}}) \right) + \frac{2\beta}{k} \mathrm{KL}(\mu\|\mu_{\mathrm{ref}}) \propto \exp(-\bar{g}^{(k)})\mu_{\mathrm{ref}}$$

Doob h-transform：

$$\mathrm{d}\bar{Y}_t = (\bar{Y}_t + 2s(\bar{Y}_t, \overline{T} - t) + 2\nabla_x \log(\mathbb{E}[\exp(-\hat{g}(Y_{\overline{T}}))|Y_t = x]|_{x=\bar{Y}_t}))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t \quad \Rightarrow \quad \bar{Y}_{\overline{T}} \sim \exp(-\hat{g})\mu_{\mathrm{ref}}$$



拡散モデルのファインチューニング

$$\min_{\mu \in \mathcal{P}} \mathcal{L}(\mu) = F(\mu) + \beta \mathrm{KL}(\mu || \mu_{\mathrm{ref}})$$

E.g.:
DPO,
Bayes filtering

μ_{ref}: Pretrained diffusion model

Diffusion model

「An astronaut riding a horse in a photorealistic style」

「Teddy bears shopping for groceries in the style of ukiyo-e」

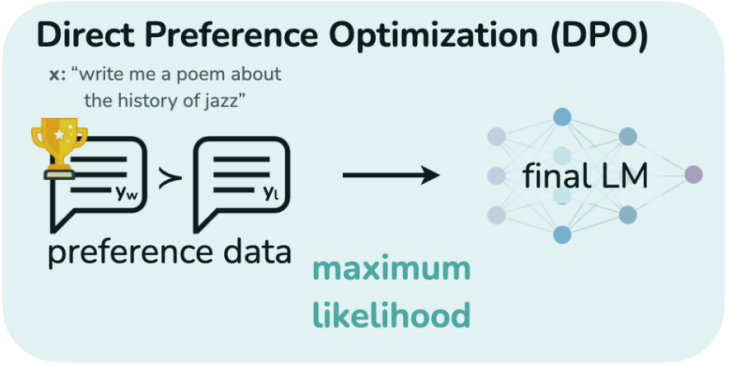
DALL·E: [Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, Ilya Sutskever: Zero-Shot Text-to-Image Generation. ICML2021.]
DALL·E2:[Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, Mark Chen: Hierarchical Text-Conditional Image Generation with CLIP Latents. arXiv:2204.06125]

SORA
(OpenAI, 2024)

Optimizing distribution

- Post training:
  e.g., Preference optimization

$$\min_{\hat{p}} \mathbb{E}_{(y_w, y_l, c)} \left[\sigma \left(\beta^{-1} \log \left(\frac{\hat{p}(y_w|c)}{p_{\mathrm{ref}}(y|c)} \right) - \beta^{-1} \log \left(\frac{\hat{p}(y_l|c)}{p_{\mathrm{ref}}(y|c)} \right) \right) \right]$$



[Rafailov et al. 2024]

- Bayesian inference
- Reinforcement learning

拡散モデルのfine-tuning

[Kawata, Oko, Nitanda, Suzuki: Direct Distributional Optimization for Provable Alignment of Diffusion Models. ICLR2025]

$$\min_{\mu\in\mathcal{P}} \mathcal{L}(\mu) = F(\mu) + \beta \mathrm{KL}(\mu||\mu_{\mathrm{ref}})$$

- μ_{ref}: 参照モデル (事前学習モデル)

双対平均化法 (Dual averaging method)

Phase 1: 最適な分布との密度比を求める.

For $k = 1, \ldots, K-1$:

$$\mu^{(k+1)} = \arg\min_{\mu\in\mathcal{P}} \frac{2}{k(k+1)} \sum_{j=1}^{k} j \left(\mathbb{E}_{\mu}\left[\frac{\delta F(\mu^{(j)})}{\delta \mu} \right] + \beta \mathrm{KL}(\mu\|\mu_{\mathrm{ref}}) \right) + \frac{2\beta}{k} \mathrm{KL}(\mu\|\mu_{\mathrm{ref}})$$

$$\propto \exp(-\bar{g}^{(k)})\mu_{\mathrm{ref}}$$

where $\bar{g}^{(k)} = \sum_{j=1}^{k} \frac{j}{\beta(k+1)(k+2)/2} \frac{\delta F(\mu^{(j)})}{\delta \mu}$

$O(1/K)$ convergence

$$\Rightarrow \quad \frac{\mathrm{d}\hat{\mu}}{\mathrm{d}\mu_{\mathrm{ref}}} \propto \exp(-\hat{g})$$

最適な分布と参照分布の**密度比**が求まる

Doob h-transform

Phase 2: 最適分布からのサンプリング $\hat{\mu} \propto \exp(-\hat{g})\,\mu_{\mathrm{ref}}$.

Doob h-Transform (Doob, 1957; Rogers & Williams, 2000)

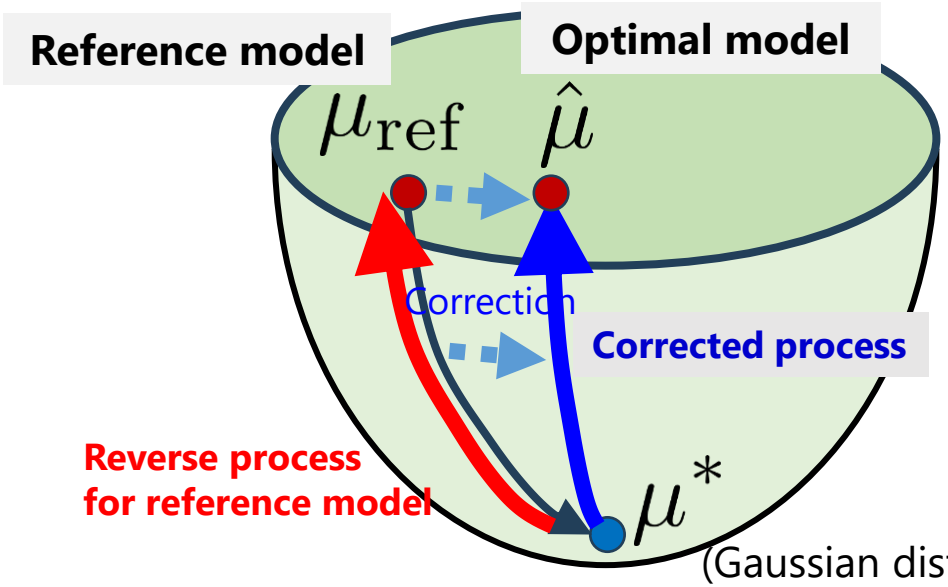
$$\mathrm{d}\bar{Y}_t = (\bar{Y}_t + 2s(\bar{Y}_t, \overline{T} - t) + 2\underbrace{\nabla_x \log(\mathbb{E}[\exp(-\hat{g}(Y_{\bar{T}}))|Y_t = x]|_{x=\bar{Y}_t})}_{\text{修正項}})\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$

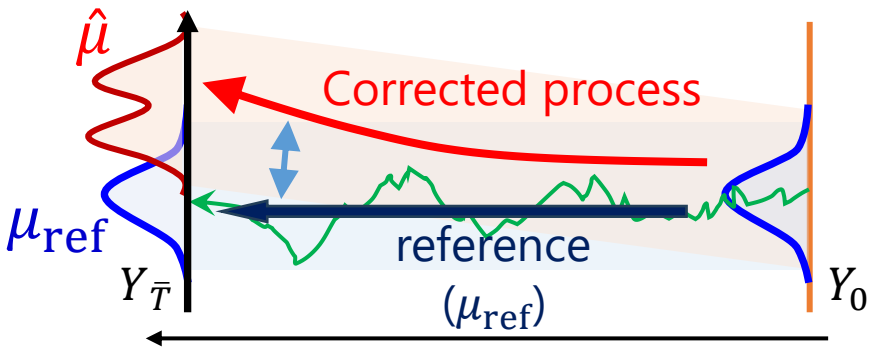
$$\bar{Y}_{\bar{T}} \sim \exp(-\hat{g})\mu_{\mathrm{ref}} = \hat{\mu}$$

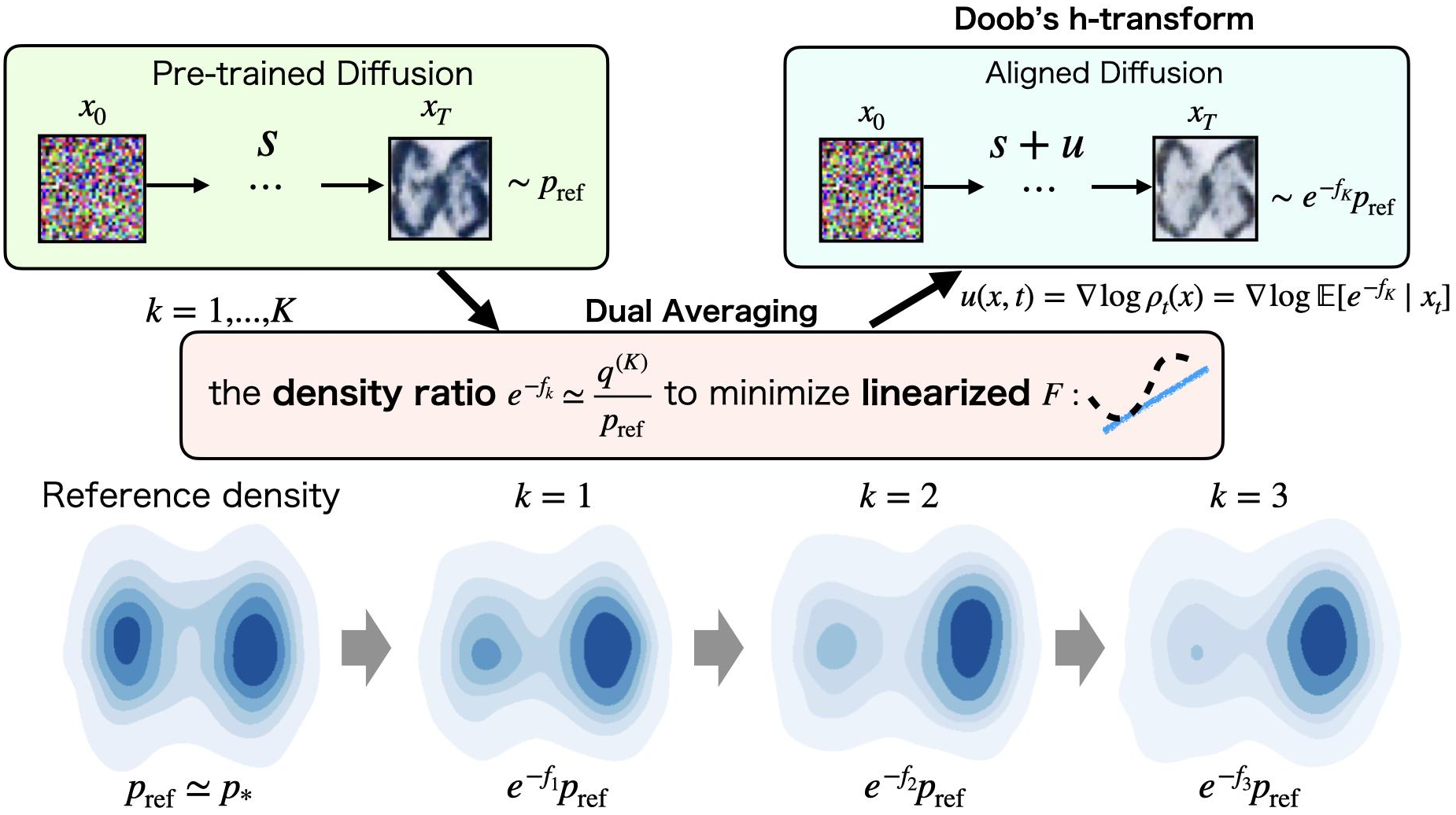
See also Vargas, Grathwohl, Doucet (2023) & Heng, De Bortoli, Doucet (2024), Uehara et al. (2024) for more details.

Reverse process of μ_{ref}:

$$\mathrm{d}Y_t = (Y_t + 2s(Y_t, \overline{T} - t))\mathrm{d}t + \sqrt{2}\mathrm{d}B_t$$






Doob's h-transform
Pre-trained Diffusion
x_0
s
x_T
$\sim p_{\text{ref}}$
Aligned Diffusion
x_0
$s+u$
x_T
$\sim e^{-f_K} p_{\text{ref}}$
$k = 1,\ldots,K$
Dual Averaging
$u(x,t) = \nabla \log \rho_t(x) = \nabla \log \mathbb{E}[e^{-f_K} \mid x_t]$
the **density ratio** $e^{-f_k} \simeq \frac{q^{(K)}}{p_{\text{ref}}}$ to minimize **linearized** F :
Reference density
$k = 1$
$k = 2$
$k = 3$
$p_{\text{ref}} \simeq p_*$
$e^{-f_1} p_{\text{ref}}$
$e^{-f_2} p_{\text{ref}}$
$e^{-f_3} p_{\text{ref}}$

Attentionの代替手段（線形手法）

線形注意機構

$\exp(k_i^\top q_i)$ を $\phi(k_i)^\top \phi(q_i)$ で置き換え

（ϕ は非線形写像）

$$y_i = \sum_{j=1}^{m} A_{i,j} v_j \qquad A_{i,j} = \frac{\phi(k_j)^\top \phi(q_i)}{\sum_{j'} \phi(k_{j'})^\top \phi(q_i)}$$

$$y_i = \sum_j v_j \frac{\phi(k_j)^\top \phi(q_i)}{\sum_{j'} \phi(k_{j'})^\top \phi(q_i)} = \frac{\left(\sum_j v_j \phi(k_j)^\top\right) \phi(q_i)}{\left(\sum_{j'} \phi(k_{j'})^\top\right) \phi(q_i)}$$

逐次的に計算可能

再帰的に計算 ➡ 計算量 $O(L)$

状態空間モデル (SSM)

内部状態 (ベクトル) で過去情報を保存

内部状態（過去の情報を保存）

$$x_{j+1} = Ax_j + Bu_j$$
$$y_j = Cx_j + Du_j$$

A, B, C, D は学習パラメータ

$$y_j = \sum_{n=0}^{\infty} (CA^n B + D\delta_n) u_{j-n}$$

H_n フィルタ（事前計算可能）

再帰的な計算 ➡ $O(L)$

FFT による畳み込み ➡ $O(L \log L)$

線形注意： $\exp(k_j^\top q_i) = \sum_{m=0}^{\infty} \phi_m(k_j)\phi_m(q_i) \approx \sum_{m=0}^{M} \phi_m(k_j)\phi_m(q_i) = \phi(k_j)^\top \phi(q_i)$

カーネル関数の有限和近似

SMM+gating： $y_j = \psi_2(u_j) \odot \left(\sum_{n=0}^{\infty} \underbrace{h_n \cdot \psi_1(u_{j-n})}_{\to \tilde{\psi}_1(u_{j-n})} \right) \Rightarrow \sum_{n=0}^{\infty} \psi_1(u_{j-n})^\top \tilde{\psi}_2(u_j)$

SSMの表現力

Naoki Nishikawa, Taiji Suzuki: State Space Models are Comparable to Transformers in Estimating Functions with Dynamic Smoothness. ICLR2025.

- **主結果１**

既存結果: Copying タスクで 1層のSSM は Transformer を代替できない

[Jelassi et al.: Repeat After Me: Transformers are Better than State Space Models at Copying. 2024]

なぜか？⇒入力依存で重要なトークンを抽出可能する必要がある

p	1	u	7	v	4	w	7	t	9	u	▶	7

b	3	o	6	d	4	t	2	s	9	s	▶	9

本研究：多層の FNN + SSM で Transformer を代替可能

各トークンの重要度を前の層で計算すれば実は代替できる：

p	1	u	7	v	4	w	7	t	9	u	▶	？
1	3	20	159	2	2	4	3	2	3	24	…	重要度（入力依存）

e.g. 自分 or 1つ前が最後のトークンと同じ

- **主結果２**

区分的γ-平滑関数の推定において，SSMとTransformerは同じ推定誤差を達成することを証明

線形注意機構によるAttentionの近似

Naoki Nishikawa, Rei Higuchi, Taiji Suzuki: Degrees of Freedom for Linear Attention: Distilling Softmax Attention with Optimal Feature Efficiency. NeurIPS2025.

- Attentionのカーネル関数としての定式化と分解：

$$K(x,y) = \exp(x^\top y/\sqrt{d}) = \mathbb{E}_{z\sim N(0,I)}[\phi(x;z)\phi(y;z)]$$

$$\text{where } \phi(x;z) = \exp\left(\frac{z^\top x}{d^{1/4}} - \frac{\|x\|^2}{2\sqrt{d}}\right)$$

- 線形注意機構と特徴写像：

$$\hat{K}(x,y) = \varphi(x)^\top \varphi(y) \quad \text{where } \varphi\colon \mathbb{R}^d \to \mathbb{R}^M.$$

定理 (近似誤差の理論評価)

$\lambda > 0$に対し，N_λをAttentionの**統計的自由度**とする．すると，

$$M \geq \frac{4}{t} N_\lambda \log\left(\frac{32N_\lambda}{t/2}\right),$$

とすることで，Attentionは線形注意機構で以下の誤差で近似できる:

$$\|K - \hat{K}\|^2_{L^2(P_X \otimes P_X)} \leq 2t\left(\lambda^2 + \|K\|^2_{L^2(P_X\otimes P_X)}\right)$$

(ただし，特徴写像φは適切に選ぶとする)

統計的自由度: $N_\lambda = \mathrm{Tr}[\Sigma_K(\Sigma_K + \lambda I)^{-1}]$

実験結果 (線形アテンションの次元)

[Naoki Nishikawa, Rei Higuchi, Taiji Suzuki: Degrees of Freedom for Linear Attention: Distilling Softmax Attention with Optimal Feature Efficiency. NeurIPS2025]

- Models: GPT-2 (1.24×10^8 parameters) [Radford et al., 2019], Pythia-1B (1.01×10^9 parameters) [Biderman et al., 2023]
- Hyperparameters: $(C, \lambda) = (64, 10^{-4})$ for GPT-2, $(C, \lambda) = (128, 10^{-8})$ for Pythia-1B

✓ Cost C is set to be the same as the head size following prior work (e.g. [Chen et al., 2025])

Model	Cost	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16
GPT-2	64	**130**	**182**	35	44	42	65	**92**	33	28	34	46	39	-	-	-	-
Pythia-1B	128	**277**	138	**275**	132	204	167	115	142	**231**	64	96	73	40	52	34	9

Large in early and middle layers　　Small in latter layers

深いレイヤーの方が少ない特徴量でOK

(CNN, FNNと同様の現象 [Arora et al., 2018; Ravichandran et al., 2019; Suzuki et al., 2020])

Downstream taskの性能

対抗手法：Performer [Choromanski et al., 2021], DiJiang [Chen et al., 2025]

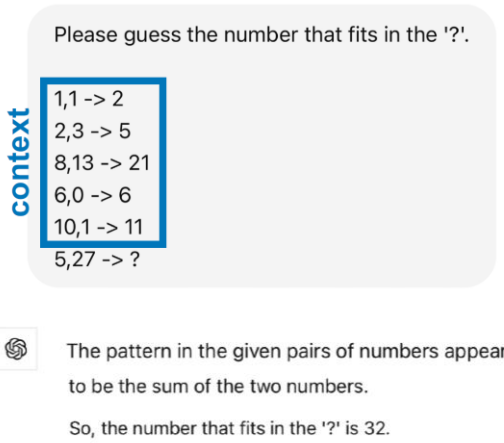
Strategy	Method	PiQA	logiQA	ARC-E	ARC-C	Winogrande	MMLU	WSC	Average
Original GPT-2		**0.5985**	0.3103	**0.3325**	0.3003	0.5122	0.2789	**0.6538**	**0.4266**
—	DiJiang	0.5065	0.2550	0.2113	0.2244	0.4846	0.2639	0.4615	0.3409
	Performer	0.5468	0.2934	0.3039	0.2747	0.4996	0.2517	0.5962	0.3952
Fix	direct	0.5832	**0.3195**	0.2921	0.2995	**0.5335**	0.2552	0.6154	0.4141
	softmax	0.5718	0.2673	0.2479	**0.3029**	0.5020	0.2634	0.6154	0.3958
	L^2	0.5822	**0.3195**	0.2483	0.2773	0.5107	0.2520	0.5962	0.3980
DoF	direct	0.5669	0.3011	**0.3241**	**0.3012**	**0.5280**	0.2712	**0.6346**	0.4182
	softmax	0.5751	0.3026	0.3224	0.2995	**0.5328**	0.2564	**0.6442**	**0.4190**
	L^2	0.5664	0.3088	0.2736	0.2824	0.4972	0.2608	0.5865	0.3965
DoF + Clip	direct	**0.5892**	**0.3164**	0.3136	0.2952	0.5075	**0.2993**	**0.6346**	**0.4223**
	softmax	**0.5860**	0.3026	**0.3401**	0.2816	0.4996	**0.2832**	**0.6346**	0.4182
	L^2	0.5822	**0.3164**	0.2942	**0.3063**	0.5091	**0.2799**	0.5673	0.4079

(特徴次元を固定した方法)

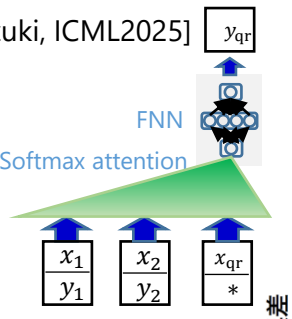
(特徴次元を層ごとに選んだ方法 [提案法])

テスト時推論の理論

文脈内学習 [Oko, Song, Suzuki, NeurIPS2024; Nishikawa, Song, Oko, Wu, Suzuki, ICML2025]

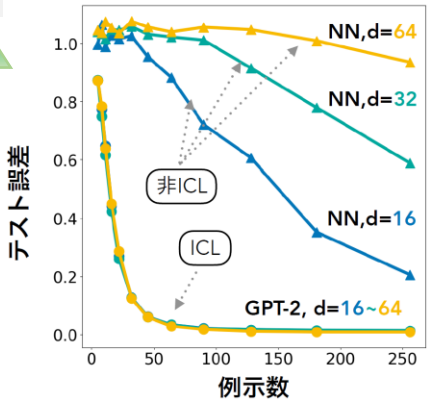


Transformerによる文脈内学習のメカニズムは？



d: 入力の次元，r: 内在的次元

	事前学習なし		文脈内学習	
手法	Kernel	NN	線形注意	非線形注意
サンプル複雑度	d^P	$d^{\Theta(\mathrm{ge}(\sigma^*))}$	r^{4P}	$r^{3\mathrm{ge}(\sigma^*)/2}$

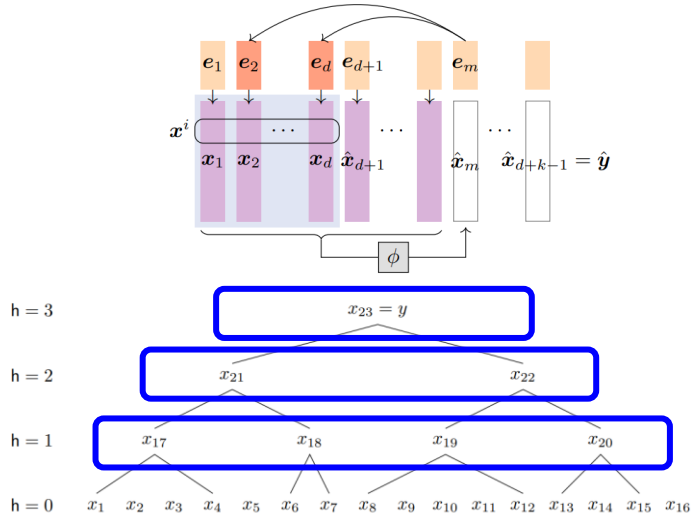


- Transformerはデータからの情報抽出の方法を事前学習によって獲得
- 勾配法による最適化の保証付き，ミニマックス最適性

思考連鎖 [Kim, Suzuki, **ICLR2025, oral**]

問題：k-パリティ問題

	通常学習	思考連鎖
サンプル/計算複雑度	$\Omega(d^{k-1})$	$O(d^{2+\epsilon})$



：中間結果→Transformerに出力させるよう訓練

<u>思考連鎖によって学習の効率が大幅に改善</u>

文献とまとめ

深層学習の適応力: ミニマックス最適性

- Besov空間 & 混合平滑Besov空間 (Suzuki, 2019)

Benefit of feature learning

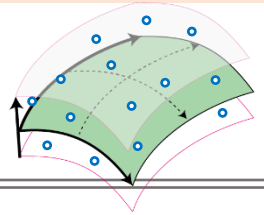
DNNは次元の呪いを回避する

- 非等方Besov空間 (Suzuki & Nitanda, 2021); γ-平滑関数空間 (Okumoto & Suzuki, 2021)

深層基盤モデルの最適性

- 拡散モデル: Oko,Akiyama&Suzuki (2023)
- Transformer: Takakura&Suzuki (2023)

$$\mathbb{E}_{D_n}\left[\mathrm{TV}(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s}{2s+d}} \log^9(n),$$
$$\mathbb{E}_{D_n}\left[\mathrm{W}_1(\hat{Y}, X_0)\right] \lesssim n^{-\frac{s+1-\delta}{2s+d'}}$$

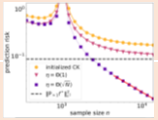


特徴学習の最適化理論

平均場ランジュバン動力学の収束

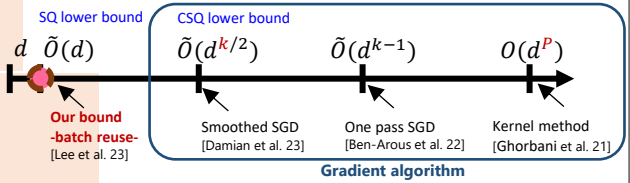
- 線形収束: Nitanda, Wu, Suzuki (2022)
- Propagation-of-chaos: Suzuki, Nitanda, Wu (2023)

$$\mathcal{L}(\mu_t) - \mathcal{L}(\mu^*) \leq \exp(-2(\lambda_2/\alpha)t)(\mathcal{L}(\mu_0) - \mathcal{L}(\mu^*)),$$
$$\lambda_2 \mathrm{KL}(\mu||\mu^*) \leq \mathcal{L}(\mu) - \mathcal{L}(\mu^*) \leq \lambda_2 \mathrm{KL}(\mu||p_\mu).$$



高次元学習問題における特徴学習

- 一回更新勾配法: Ba, Erdogdu, Suzuki, Wang, Wu, Yang (2022)
- (近似的)情報理論的下限の達成: Lee, Oko, Suzuki, Wu (2024)



内在的次元を用いた深層学習の汎化誤差

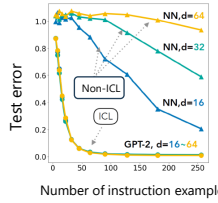
- 圧縮型バウンド: Suzuki, Abe, Nishimura (2020)

モデル圧縮への応用

- **Spectral pruning:** Suzuki et al. (2020)

テスト時推論の理論

- **文脈内学習**: Oko, Song, Suzuki, Wu (2024); Kim, Nakamaki, Suzuki (2025)
- **思考連鎖**: Kim&Suzuki (2025); Kim, Wu, Lee, Suzuki (2025)



連続最適化チーム

求解困難な数理最適化問題に対して理論保証付き効率的解法を提案

1. AI for Scienceで使われる非凸スパース最適化法の学習高速化
   - 問題点：スパース正則化関数，深層学習の非平滑活性化関数などの非凸非平滑性が理論的・実際的に効率的な最適化法の構築を阻む．
   - 2018年に近接DC最適化法(pDCA)を提案し，L0スパース正則化問題に対する効率的解法として定着．
2. 敵対的攻撃対策やハイパーパラメータ調整に役立つ意思決定多段階化
   - 問題点：最も単純な2段階最適化問題に対して超勾配を用いた効率的解法が提案されたが，他ケースについては依然として求解困難．
   - 非平滑2段階問題，リーマン2段階問題，多段階最適化問題へと拡張．いずれも，収束保証のある勾配ベースの効率的解法を初めて提案．
   - 連続最適化分野最大国際会議ICCOPT2022（3年おき開催）のsemi-plenary講演者として本成果報告．
3. ランダム行列理論に基づく巨大学習モデルの極小化
   - 問題点：特別な構造をもたない高次元最適化問題の解法研究は長らく大きな進展がない．
   - 2019年頃から，通常の非線形最適化法にランダム射影技法を組み込む解法構築が始まる．我々もこれまで8本の論文あり．
   - 本成果をうけて，数理最適化分野における最大の国際会議ISMP2027（3年おき開催）の「Random Methods for Continuous Optimization」のStream Organizerに選出

連続最適化チーム

求解困難な数理最適化問題に対して理論保証付き効率的解法を提案

1. 非凸非平滑最適化問題
   - 問題点：スパース正則化関数，深層学習の非平滑活性化関数など，非平滑性はよく現れるが，この性質は理論的・実際的に効率的な最適化法の構築を阻む．
   - 2018年に近接DC最適化法(pDCA)を提案し，L0スパース正則化問題に対する効率的解法として定着．
2. 多段階最適化問題
   - 問題点：Franceschi et al. [ICML '17] の研究により，最も単純な2段階最適化問題に対して超勾配(hypergradient) を用いた効率的解法が提案されたが，他ケースについては依然として求解困難．
   - 超勾配に基づく効率的解法を非平滑2段階問題，リーマン2段階問題，多段階最適化問題へと拡張．いずれも，収束保証のある勾配ベースの効率的解法を初めて提案．
   - 連続最適化分野最大国際会議ICCOPT2022（3年おき開催）のsemi-plenary講演者として本成果報告．
3. 高次元最適化問題
   - 問題点：高サンプルからなる確率最適化問題については研究が進展している一方，特別な構造をもたない高次元最適化問題の解法研究は長らく大きな進展がない．
   - 2019年頃から，通常の非線形最適化法にランダム射影技法を組み込む解法構築が始まる．我々もこれまで8本の論文あり．
   - 本成果をうけて，数理最適化分野における最大の国際会議ISMP2027（3年おき開催）の「Random Methods for Continuous Optimization」のStream Organizerに選出

成果1: 非凸スパース最適化の解法（pDCA）

非凸スパース最適化

$$\min_{\boldsymbol{w}\in S} \quad f(\boldsymbol{w}) \qquad \text{subj.to} \quad \|\boldsymbol{w}\|_0 \le K$$

$$\min_{\boldsymbol{W}\in S} \quad f(\boldsymbol{W}) \qquad \text{subj.to} \quad \mathrm{rank}(\boldsymbol{W}) \le K$$

応用例：スパース回帰問題
スパースポートフォリオ問題など

応用例：低ランク行列補完問題など

- 今までは，非凸スパース最適化問題に対して，非凸最適化問題を解かないですむように問題を緩和
  - ✓ 非凸のまま扱う
  - ✓ 問題の特徴(Difference of Convex Functions) を生かした解法で，最適性の必要条件を満たす解（停留解）を高速に求める

Goto, Takeda, Tono [MathProg' 18]

非凸非平滑最適化解法（SDCAM）へと拡張

$$\min_{\boldsymbol{w}\in C} f(\boldsymbol{w}) \quad \text{s.t.} \ \|\boldsymbol{w}\|_0 \le K$$ 非凸スパース最適化

$$\Leftrightarrow \min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + \delta_{\{\boldsymbol{w}:\|\boldsymbol{w}\|_0\le K\}}(\boldsymbol{w}) + \delta_C(\boldsymbol{w})$$

$$\Rightarrow \min_{\boldsymbol{w}\in\mathbb{R}^n} f(\boldsymbol{w}) + \sum_{i=1}^{m} g_i(\boldsymbol{w})$$ より一般的な非凸非平滑最適化

- 非凸スパース最適化問題に対する解法に工夫を加え，より一般的な非凸非平滑最適化問題に対する解法SDCAMの構築 Liu, Pong, Takeda [MathProg ’19]
- 損失関数（期待値）計算をランダムサンプリングで置き換える確率的SDCAMの開発 Metel, Takeda [ICML ’19]

をはじめとする4本の研究成果に加えて．．．

- 最近，SDCAMのシングルループ化に成功 Zhang, Marumo, Pong, Takeda [投稿中, ’25]

成果2: 多段階最適化

これまでの
2レベル最適化問題

$$\min_{x\in X} \; F(x) := f(x, y^*(x))$$
$$\text{s.t. } \; y^*(x) = \arg\min_{y'} g(x, y')$$

これまでの仮定: 下位レベル問題に(1)制約式がない, (2)目的関数gが微分可能な強凸であること をなくして, より広いクラスの2レベル最適化問題を扱えるようにしたい

研究2-1

「(1)制約式がない」の仮定を外す

研究2-2

「(2) gが微分可能な強凸であること」の仮定を外す

研究2-3

$$\min_{x\in\mathcal{M}_x} \; F(x) := f(x, y^*(x))$$
$$\text{s.t. } y^*(x) = \arg\min_{y\in\mathcal{M}_y} g(x, y)$$

- $\mathcal{M}_x, \mathcal{M}_y$: Riemann多様体
- $f, g: \mathcal{M}_x \times \mathcal{M}_y \to \mathbb{R}$: 微分可能, 勾配ベクトルがリプシッツ連続
- $g(x, y)$: yに関して測地的強凸

$$\min_{x,y,z} f(x, y, z)$$
$$\text{s.t. } y \in \arg\min_{y',z'} g(x, y', z')$$
$$\text{s.t. } z' \in \arg\min_{z''} h(x, y', z'')$$

下位レベル問題の制約としてさらに最適化問題が現れる, 3レベル最適化問題

$$\min_{\omega^*_\lambda, \lambda} \; f(\omega^*_\lambda)$$
$$\text{s.t } \; \omega^*_\lambda \in \operatorname*{argmin}_{\omega\in\Re^n} G(\omega, \bar{\lambda}) + \lambda_1 R_1(\omega)$$
$$(\lambda_1, \bar{\lambda}) \in \Omega_\epsilon \subset \Re^r,$$

$$G(\omega, \bar{\lambda}) := g(\omega) + \bar{\lambda}^T \bar{R}(w)$$
$$\bar{\lambda} := (\lambda_2, \ldots, \lambda_r)^T$$
$$\bar{R}(w) := (R_2(w), \ldots, R_r(w))^T$$
$$R_1(\omega) := \sum_{i=1}^{n} \psi(|\omega_i|^p) \quad (0 < p \le 1)$$
$$\Omega_\epsilon := \{(\lambda_1, \bar{\lambda}) \in \Re \times \Re^{r-1} : \lambda_1 \ge \epsilon, \bar{\lambda} \ge 0\}$$

非凸
非平滑

成果2-1：Riemann多様体上の2段階最適化

これまでの問題

$$\min_{x\in X} \; F(x) := f(x, y^*(x))$$
$$\text{s.t. } y^*(x) = \arg\min_{y'} g(x, y')$$

→

$$\min_{x\in \mathcal{M}_x} \; F(x) := f(x, y^*(x))$$
$$\text{s.t. } y^*(x) = \arg\min_{y\in \mathcal{M}_y} g(x, y)$$

[Han, Mishra, Jawanpuria, Takeda, NeurIPS, 2024]

- これまでの超勾配（目的関数Fの勾配）を用いた解法では，下位レベル問題に制約式を課すことができなかった．
- Riemann多様体上での2レベル最適化問題を考えることにより，変数x, yの動く空間を正定値錐，Stiefel多様体（直交制約つきの行列空間），2重確率行列などに制限することができる．
- fのRiemann超勾配（FのRiemann勾配）の推定を初めて行なった．具体的にはHesse作用素の逆作用素・共軛勾配法・Neumann級数・自動微分を利用した推定方法を提案し，推定誤差の分析をするとともに，数値実験で比較を行った．

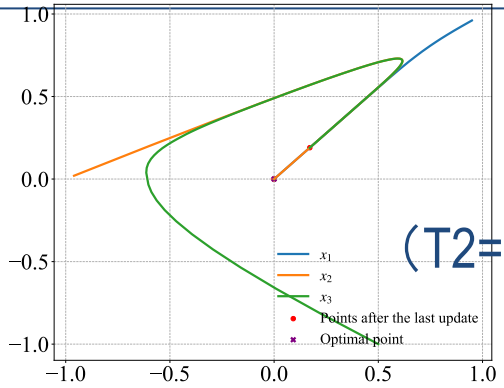
研究2-2：多レベル最適化問題の収束性保証つき解法研究

（NeurIPS 2021）

3レベル最適化問題

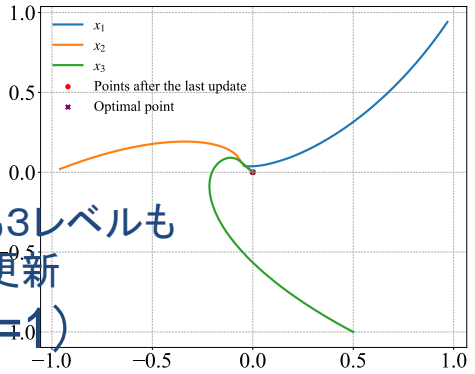
$$\min_{x,y,z} f(x,y,z)$$
$$\text{s.t. } y \in \arg\min_{y',z'} g(x,y',z')$$
$$\text{s.t. } z' \in \arg\min_{z''} h(x,y',z'')$$

- 提案解法は，結果的には，2レベル最適化問題に対する既存解法（Franceschi等，ICML'17）を多レベル最適化問題に対して愚直に拡張した，とみなせる．
- 3レベル目を最急降下法をT反復にさせてzを更新したら，2レベル目の変数yを1反復更新する，また3レベルに戻り…を繰り返して，2レベル目の変数yをT回更新させたら，ようやく1レベル目の変数xを1反復更新，といった調子．
- Tを無限回繰り返せば，提案手法は多レベル最適化問題の停留点が得られる．
- 実際には，Tを多く繰り返すのは無駄が多い（特にアルゴリズム序盤）．



各レベル1変数ずつのtoy problem

（T2=T3=10）



2レベルも3レベルも1反復の更新

（T2=T3=1）

研究2-3：微分不可能な2レベル最適化問題の最適性保証つき解法研究　1/2

問題例)

$$\min_{\lambda} \|y_{val} - X_{val} w_\lambda\|^2$$

下位レベルの最適化問題

$$\text{s.t. } w_\lambda \in \arg\min_{w} \|y_{tr} - X_{tr} w\|^2 + \lambda \|w\|_1$$

微分不可能

これまでの2レベル最適化研究

- ハイパーパラメータ学習問題はしばしば2レベル最適化問題として定式化される．L1ノルムなど微分不可能な正則化を想定する場合には，下位の最適化問題が微分不可能な問題になる．
- これまでは上位，下位の最適化問題ともに，微分可能性が仮定された解法のみ研究されてきた．その場合，下位レベルの目的関数を微分して0に置いて1レベル問題にするのが，これまでの手法である．
- 一方，下位レベル問題が微分不可能になると，(1)下位問題の最適解集合を扱わなければならない，(2)最適性条件が「目的関数を劣微分してその集合に0が含まれているか」となり，1レベル問題にして求解するのが難しい，といった問題が生じる．

研究2-3：微分不可能な2レベル最適化問題の最適性保証つき解法研究　2/2

$$\min_{\lambda} f_{\mathrm{val}}(w_\lambda)$$
$$\text{s.t. } w_\lambda \in \arg\min_{w} f_{\mathrm{tr}}(w) + \lambda \sum_{i=1}^{n} \psi(|w_i|^p)$$

微分不可能

$(0 < p \leq 1)$

- 微分不可能性のために，下位の問題は劣微分（集合）を使った表現となり，1レベル最適化問題は解く術がない．
- これまで，L1正則化ハイパーパラメータ学習問題の停留解を保証する解法は提案されていない．
- 今回，停留解のepsilon-近似解を出力する解法を提案した．2レベル最適化問題の”epsilon-近似解”という概念を，最適性必要条件（2段階KKT条件）を導くことで初めて定義し，2段階KKT条件を満たす解，つまり，epsilon-近似解が得られるような解法を構築した．また，その後，よりタイトな条件を導くことに成功し，さらに対象とする問題の一般化にも成功した．

Okuno, Takeda, Kawana, Watanabe (JMLR ’21)
Alcantara, Nguyen, Okuno, Takeda, Chen (Mathematical Programming, 2025)
Alcantara, Takeda (under review)

成果3： ランダム射影を利用した精度保証付き部分空間最適化法

- ランダム射影行列は扱うデータサイズを小さくすることを目的に機械学習分野で使われているものの，これまで最適化問題の変数サイズを小さくする目的の研究はこれまでほとんどなかった．最近，D’Ambrosio, et. al.（Mathematical Programming，‘20)では，ランダム射影行列を用いて最適化問題サイズを縮小し誤差評価をする方法，また，Gower et al.（NeurIPS‘19）らにより，ランダム射影行列を用いてニュートン法のヘッセ行列の逆行列計算を軽減する手法が提案された．この流れを受けて，2つの研究を行った．いずれも大規模非凸最適化問題に対して規模の小さい凸緩和問題を１回ないし，繰り返し解くことを提案し，理論保証を与えている．
- （1）大規模最適化問題（非凸2次最適化問題，LP，SDPなど）に対して，ランダム射影を用いて小さい規模（変数サイズk）の非凸2次最適化問題を構築し，さらに凸緩和した問題を解くことを提案．その際，kと「元問題の最適値との乖離」，さらに「凸緩和することによる乖離」を評価して近似精度を導出．

  T. Fuji, P.L. Poirion, A. Takeda, （SIAM Journal on Optimization, 2022)含む，論文3本
- (2)一般的な非凸最適化問題に対する反復解法：正則化付きニュートン法の提案．毎反復で子問題として凸2次最適化問題が解かれている．最初の研究成果と同様に，ランダム射影を用いて，その子問題の代わりに小さい規模（変数サイズk）の凸2次最適化問題を解く解法を提案し，収束スピードの評価を行った．

  T. Fuji, P.L. Poirion, A. Takeda, （Open Journal of Mathematical Optimization, 2025）や
  R. Higuchi, P.L. Poirion, A. Takeda (ICLR 2025, spotlight) を含む，論文5本

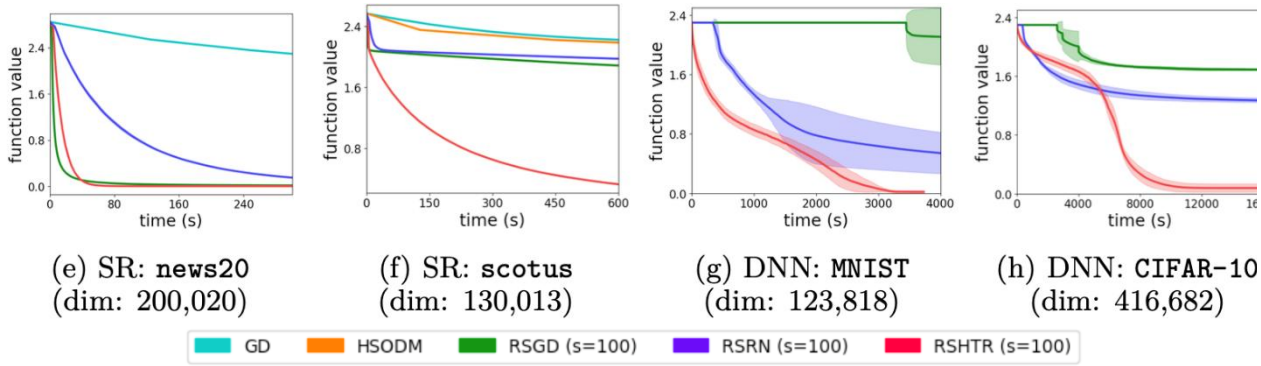
Recent work on Random Subspace Algorithms

Higuchi-Poirion-Takeda (ICLR 2025, spotlight)

Best theoretical performance among all random subspace methods

	Underlying algo.	Subprob. cost/iter	Global	Local	SOSP	Feas.
Roberts & Royer (2023)	Direct search	Multi. line-search	$O(\varepsilon^{-2})$			✓
Dzahini & Wild (2024)	Zeroth order	Finite diff. grad.	$O(\varepsilon^{-2})$			✓
Kozak et al. (2023)	Zeroth order	Finite diff. grad.	$O(\varepsilon^{-2})$	1		✓
Kozak et al. (2021)	Grad. descent	Gradient	$O(\varepsilon^{-2})$	1		✓
Cartis et al. (2020)	Gauss-Newton	Cond.quad.prog. (QP)	$O(\varepsilon^{-2})$			✓
Shao (2022)	Cubic Newton	Cond. cubic.reg.QP	$O(\varepsilon^{-3/2})$		✓	
Zhao et al. (2024)	Cubic Newton	Cubic.reg.QP	$O(\varepsilon^{-3/2})$		✓	✓
Fuji et al. (2022)	Reg. Newton	Solve eq.	$O(\varepsilon^{-2})$	$1+^{\dagger}$		✓
Ours	Trust Region	Min eigenvalue	$O(\varepsilon^{-3/2})$	2^{*}	✓	✓

Practical performance



(e) SR: `news20` (dim: 200,020) (f) SR: `scotus` (dim: 130,013) (g) DNN: `MNIST` (dim: 123,818) (h) DNN: `CIFAR-10` (dim: 416,682)