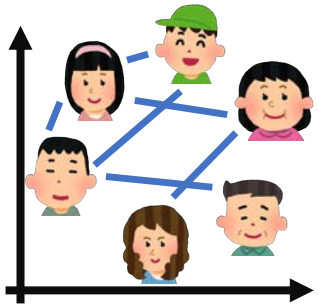
京都大学情報学研究科システム科学専攻
数理システム論分野　シー５(a)
下平・本多研究室

研究トピックの紹介

- **グラフ埋め込みの理論**
- **グラフ埋め込みの応用**
- **分散表現の論理演算**
- **マルチスケール k-近傍法**
- **複雑ネットワークの統計推測**
- **ブートストラップと選択的推測**
- **高次元小標本の統計学**
- **バンディット問題**
- **強化学習を用いたバンディットアルゴリズム**
- **ベイズ最適化**

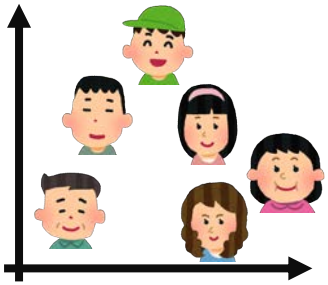
座標とリンク

（リンクだけでも良い）



ニューラルネット

座標

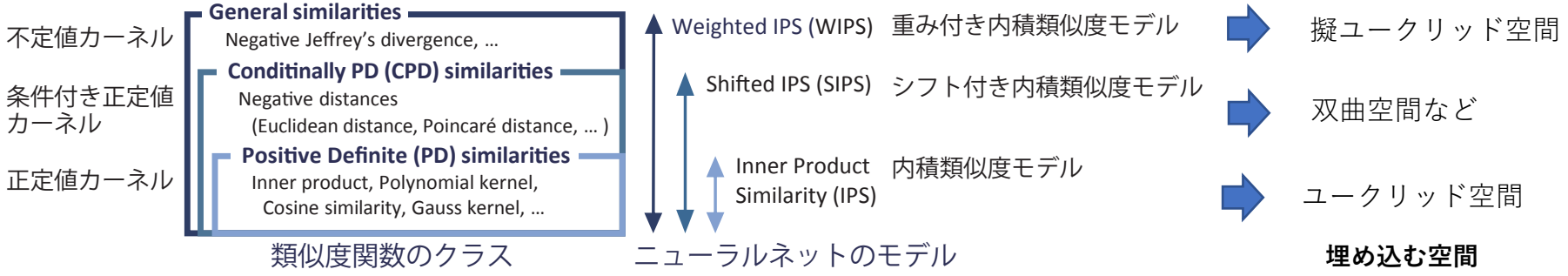


グラフのリンク構造をなるべく保存するようにノードの座標を計算する．統計学の多変量解析の拡張になっていて，機械学習で応用がたくさんある．

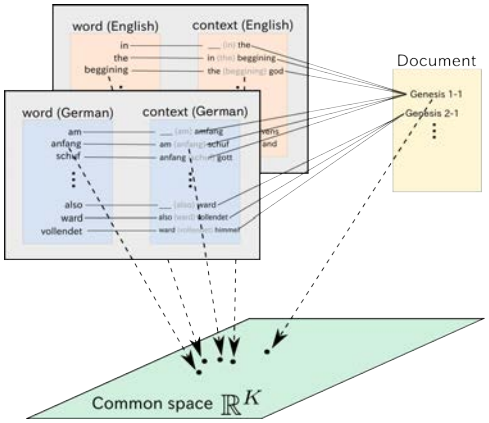
- 一般にノードの類似度をベクトルの内積やコサイン類似度で測る
- 類似度関数を変えたら，もっと性能あがるのでは？

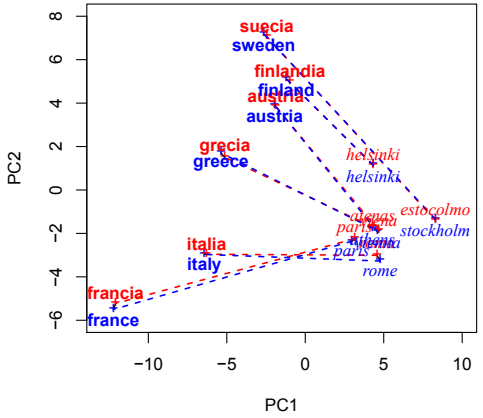
- 内積だけでも任意の正定値カーネルが表現できる！
- シフト項をいれるだけで，双曲空間への埋め込みが表現できる！
- 重み付き内積モデルにすれば，もっと表現できる！

エッシャーの絵は双曲幾何の例
（ポアンカレ埋め込み）

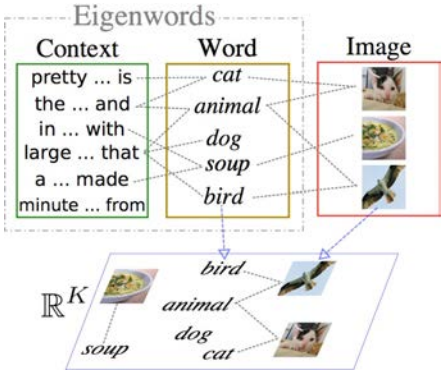


グラフ埋め込みの理論





- 対訳コーパスのリンクだけから単語埋め込み
- 辞書を教えなくても，２言語の翻訳単語の座標は近くなる



- 文書データ，ソーシャルメディアの写真とタグから，単語と画像を同時埋め込み
- 画像と単語の相互検索や，意味の加減算による検索ができる

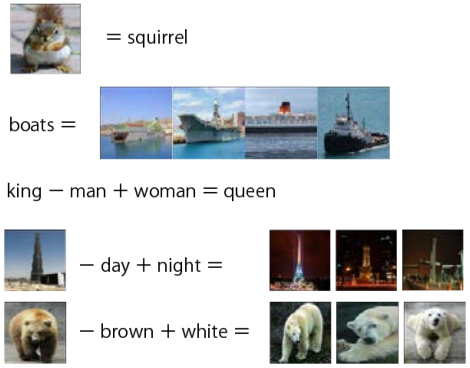



図3. Flickrデータの可視化

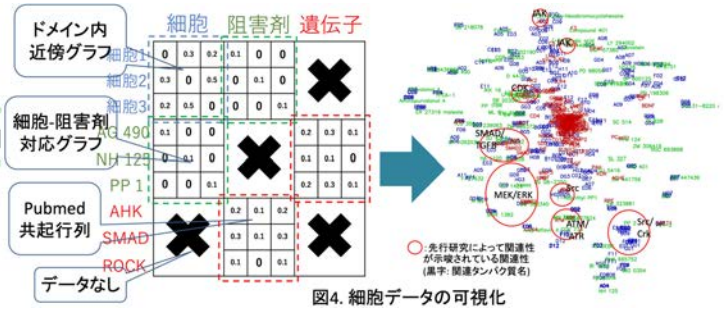


図4. 細胞データの可視化

- 文書，写真，遺伝子データなど混在した関連性データを扱える
- t-SNEなどの可視化を拡張

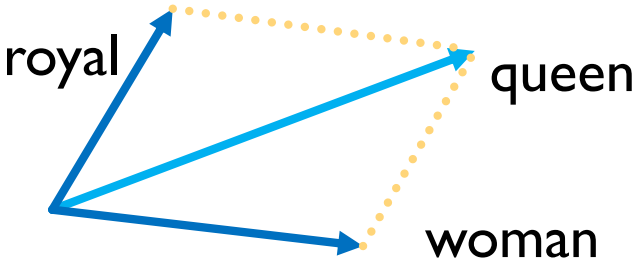
グラフ埋め込みの応用

単語埋め込みでは，文書データから教師なし学習によって「単語の意味」の分散表現（単語ベクトル）を計算する．ベクトルの加減算で類推などの「意味の計算」ができることが知られている（例：king - man + woman = queen）．**将来，高度な思考を実現するには，複雑な意味の計算が必要になる**．その基礎となる**加法構成性**では，ベクトルの和が構成要素の**ANDの意味**に相当する．

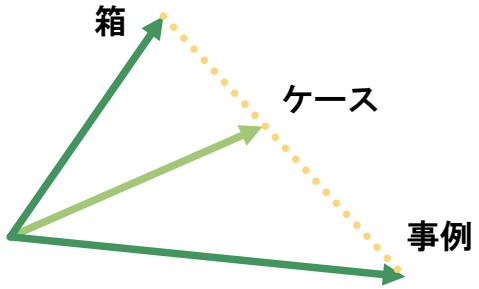
- それでは，**ORやNOTはどのように計算するのか？**
- 単語頻度による重み付き平均が**ORの意味**に相当
- NOTは むずかしい(母の反対は娘か父か？)
- 対象となる単語集合に応じて「条件付き埋め込み」の概念を導入し，ベクトルの負方向が**NOTの意味**に相当することを明らかにした

queen = royal + woman
king = royal + man

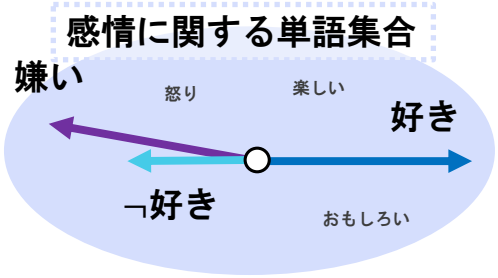
加法構成性の例．これらからroyalを消去すると， 類推の計算（king - man + woman = queen）が得られる



AND（加法構成性）



OR（ベクトルの重心）



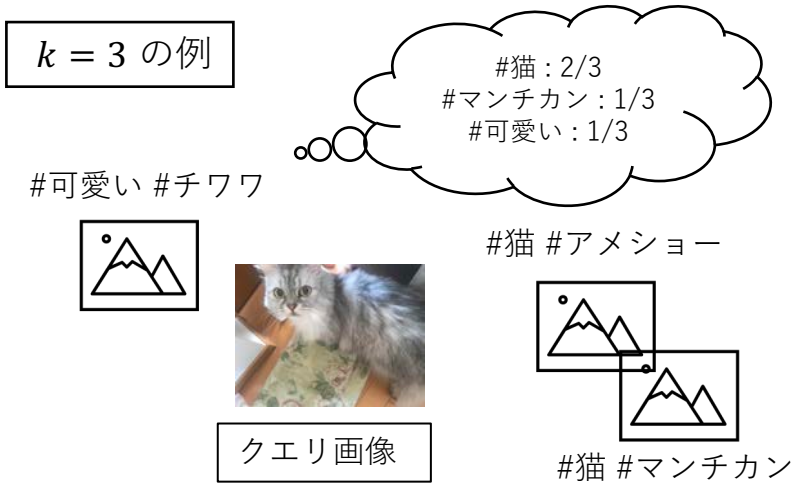
NOT（集合内で負方向）

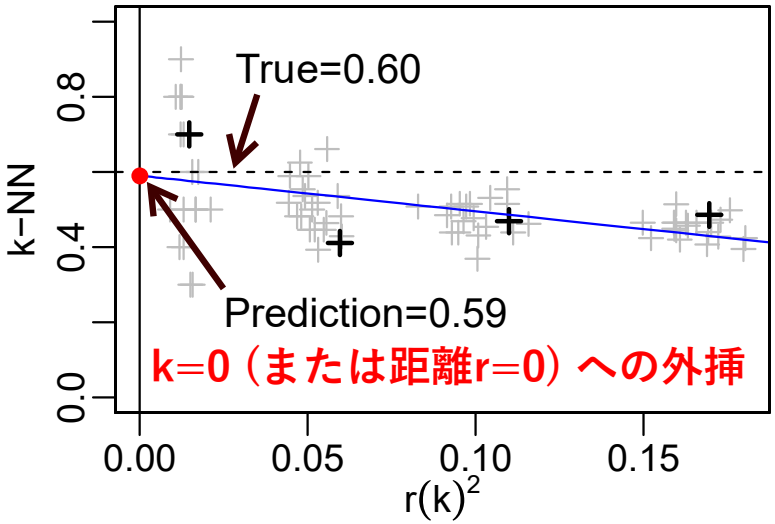
内藤，横井，下平（2021）「単語埋め込みによる論理演算」
言語処理学会第27回年次大会(NLP2021)において委員特別賞

分散表現の論理演算

k-近傍法は最も単純な判別，分類の手法．たとえば，クエリ画像に特徴量ベクトルが最も近い画像をk個選び，そのラベルの平均値を出力する．**小さいkはバイアスが小さく，大きいkは分散が小さくなり，両者の影響はトレードオフの関係**にある．どうやって最適なkを選ぶかが問題になっていた．

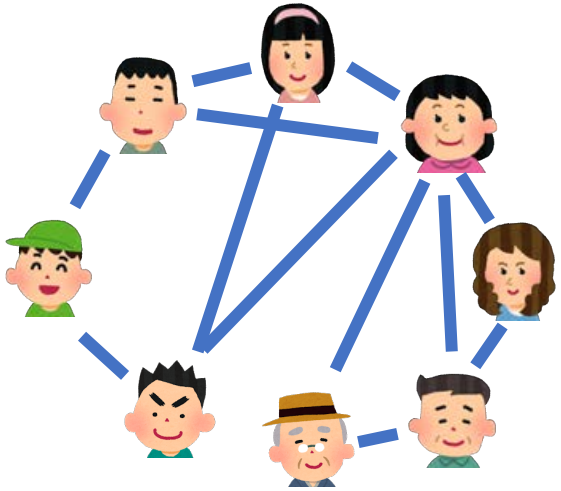
- 理論上はk=0とすればバイアスがゼロになるハズ（分散は無限になってしまう）
- **複数のkの値でk-近傍法を実行し，それをk=0に外挿すればよいのでは？**
- 架空の「0-近傍法」を計算するマルチスケールk近傍法を提案
- これが収束レートに関して最適解であることを証明
- 外挿法の工夫によって改良し，Flickr画像のタグ推定に応用





Okuno, Shimodaira (2020) “Extrapolation Towards Imaginary 0-Nearest Neighbour and Its Improved Convergence Rate” (NeurIPS 2020)
田中，奥野，下平 (2021)「マルチスケールk-近傍法による画像のExtreme Multi-Label分類」(JSAI2021)
Cao，田中，奥野，下平 (2021)「マルチスケールk-近傍法における回帰関数および損失関数の検討」(JSAI2021)

マルチスケール k-近傍法



- 成長するネットワーク（グラフ）の確率モデル
- 新しいリンクの確率が高いのは？

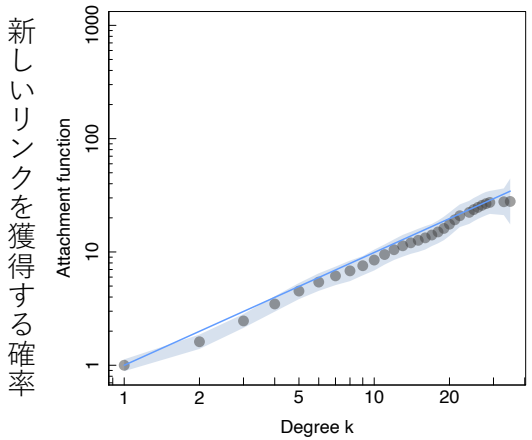
1. リンク数の多い人

（実は新人なのにすでに２人）

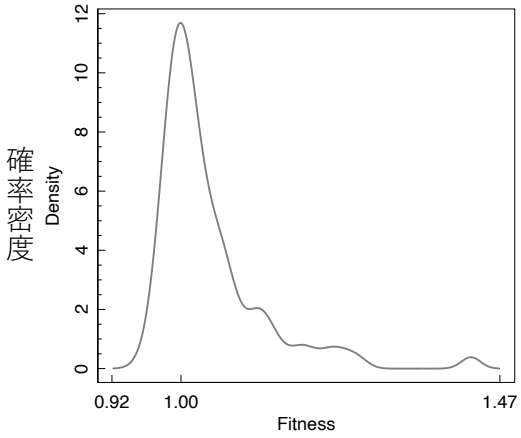
2. ポテンシャルの高い人

3. 共通する友達の多いペア

これらの効果を分離して推定する統計手法とソフトウエア(PAFit, FoFaF)の開発



現在のリンク数



新しいリンクを獲得する潜在的能力

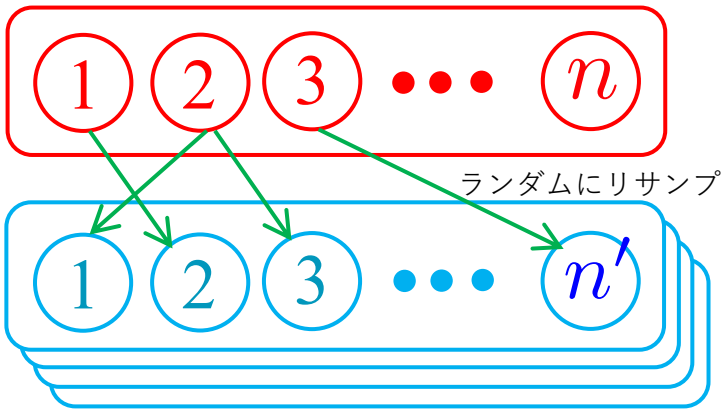
Rank	Estimated fitness	Name
1	1.42	BARABASI, A
2	1.35	NEWMAN, M
3	1.26	JEONG, H
4	1.25	LATORA, V
5	1.24	ALON, U
6	1.23	OLTVAI, Z
7	1.23	YOUNG, M
8	1.22	WANG, B
9	1.21	SOLE, R
10	1.21	BOCCALETTI, S

「コラボする潜在能力」の高い研究者トップ10

Pham, Sheridan, Shimodaira (2020)

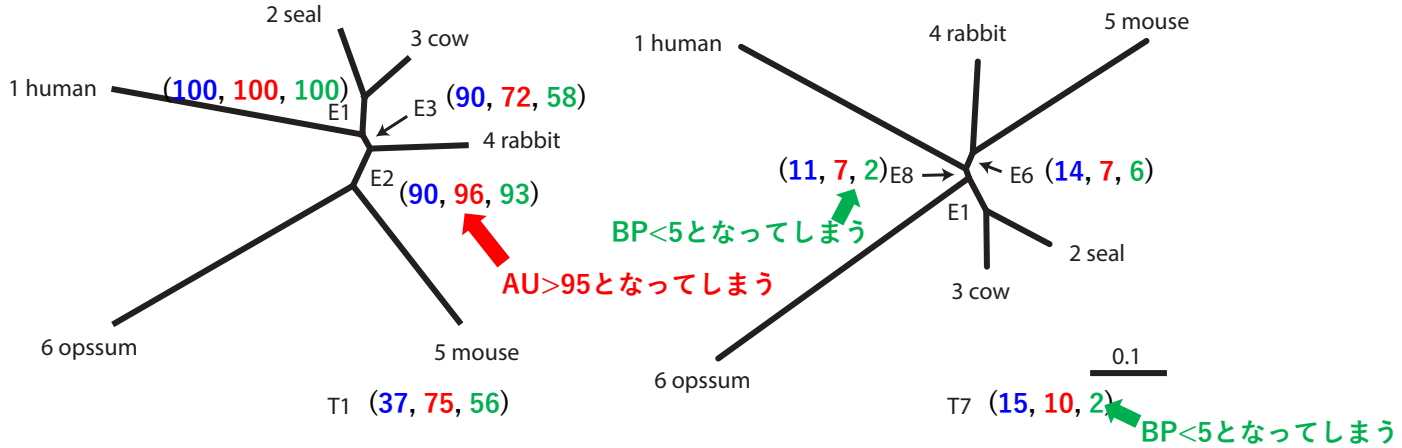
論文共著者ネットワークデータ（複雑ネットワーク分野）からの推定結果

- ブートストラップ法は，データからリサンプリング（復元抽出）
- 擬似的にたくさんデータができるので，何回もデータ解析すれば，そのバラツキもわかる
- 簡単なアルゴリズムだが，背後には確率分布の空間における距離とか曲率など高度な数理統計学



- 同じ結果が出た回数を素朴に数える(BP)は，「ニセの発見」になりやすい！
- マルチスケール・ブートストラップ法(AU)は，**n’ = -n** としてバイアスゼロ！
- じつは「膨大な仮説集合から仮説を選んでるバイアス」を直してなかった
- そのバイアスも新たに開発したばかりの選択的推測(SI)で解決できた！

SI=Selective Inference 選択的推測（提案法），AU=Approximately Unbised 近似的に不偏な検定(従来法），BP=ブートストラップ確率（ベイズ）



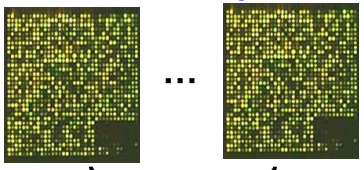
系統樹推定の例．T1はデータが支持する系統樹だが，実際にはT7が真実の系統樹と考えられる．

ブートストラップと選択的推測

- 従来の統計理論では，次元dを固定して，サンプルサイズnが大きい極限を考えてる
- ところが，nは小さいのに，dが大きいデータを分析することが増えている
- 新しい統計理論が必要！

High-dimension, low-sample-size (HDLSS) data

Microarray



n

例. 白血病や乳がん等
d (遺伝子数) ≈ 10000
n (患者数) ≈ 100

$d \gg n$

~~大標本理論~~

標本共分散行列の逆行列が不安定
各種次元の呪い

$d \to \infty,\ d/n \to \infty$
$n \to \infty$ or n: fixed

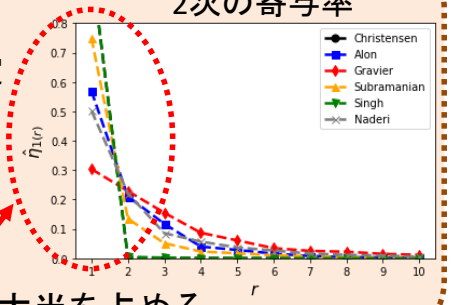
新たな理論・方法論の構築が必要

高次元小標本漸近理論

潜在空間の解析

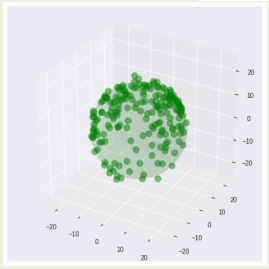
固有解析
・固有値・固有ベクトル推定
・固有値は高次元で
　不一致性を持つ
→要修正

2次の寄与率



上位の固有空間が大半を占める

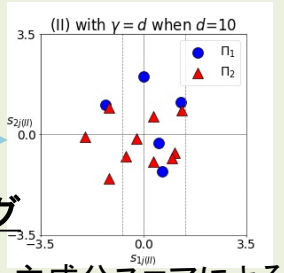
ノイズ空間の解析



球面集中現象

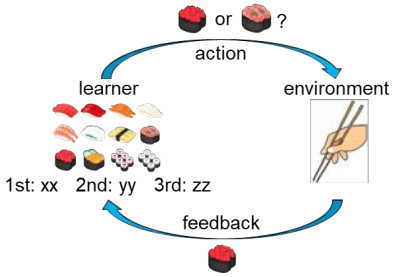
応用

判別分析
クラスタリング

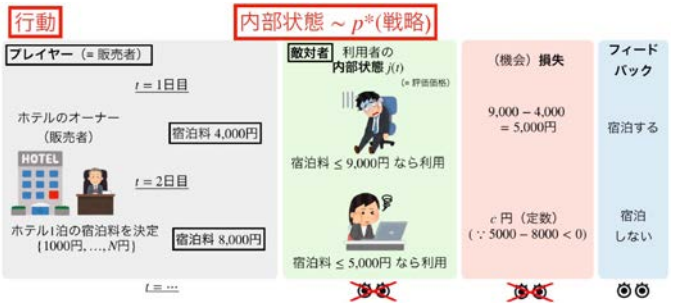


主成分スコアによる
クラスタリング

線形モデル



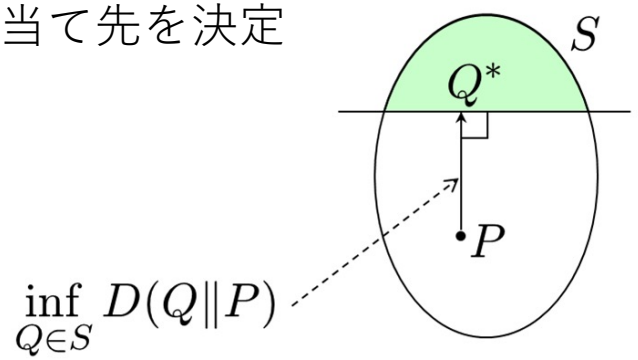
相対比較モデル

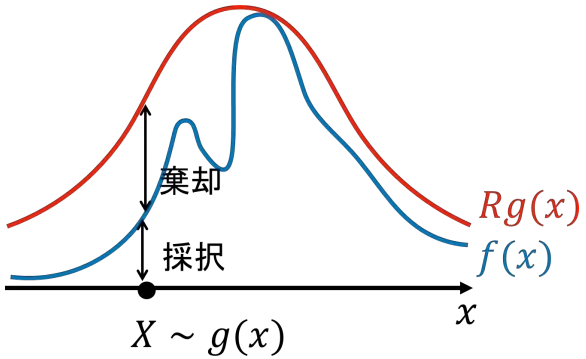


価格決定モデル

- **バンディット問題**：広告推薦や治験，価格決定問題など，事前にデータがない状態で逐次的にデータを集めながら次の割り当て先を決定
- どうやって現在の推定の不確かさを考慮する？
- 考える設定に応じて多様なモデリングや定式化

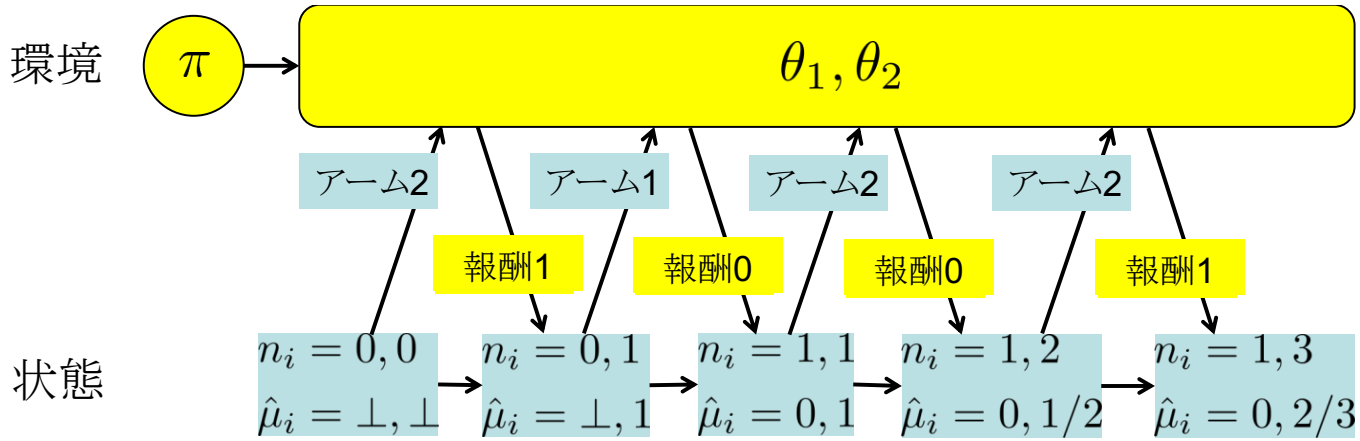
- 真の分布の候補をKL情報量を用いて抽出する！
- 理論限界を達成するアルゴリズムを
  さまざまな設定で構築可能





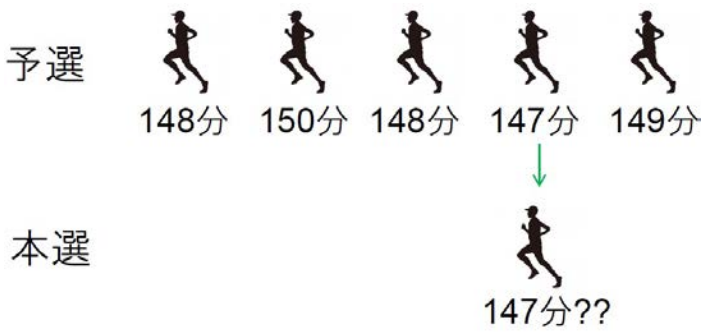
KL情報量を明示的に用いると計算が大変

ベイズ統計に基づく事後分布サンプリングを適切に用いて複雑な計算を回避！



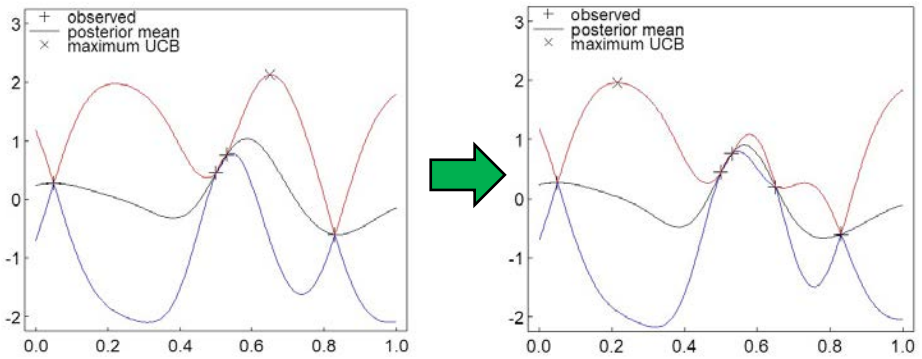
- **バンディット問題と強化学習**： バンディット問題は強化学習の特別な場合だが，実際は標準的な手法が大きく異なる
- 強化学習の手法を用いてもバンディット問題のアルゴリズムを作れる？
- 実際はランダム性の大きさ等の要因から単純なタスクを除いて学習が難しい
- 学習しやすい報酬を適切に設定して，複雑なモデル上でもうまく動くアルゴリズムを強化学習で自動設計！

- バンディットアルゴリズムを通じて得られたデータからはバイアスのない推定が難しい（選択的推定）
- 強化学習の手法を用いて推定可能性を検証する！

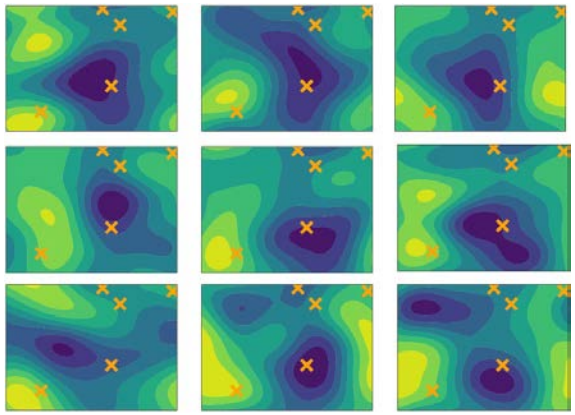


強化学習を用いたバンディットアルゴリズム

- **ベイズ最適化**： 複雑な形状の未知関数をガウス過程により表現，なるべく少ない関数評価で最大値を発見
- 各点での関数値の不確かさを定量化
- 深層学習のパラメータチューニングから物性の予測までさまざまな関数・タスクを表現できる！

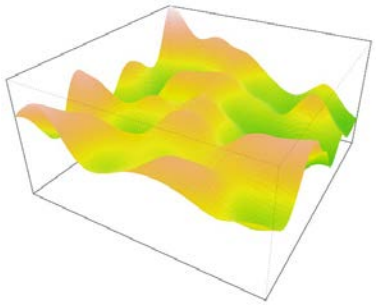


- 異なる観測コストと精度が選択できる場合，入力可能な点が事前に分からない場合，空間センシングなど時間変化がある場合等にも柔軟に対応可能

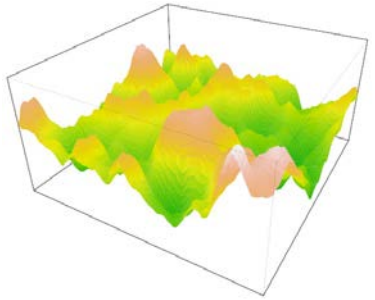


ガウス過程による電気伝導度の推定

カーネル関数を適切に設計して
関数の滑らかさをモデリング



ガウスカーネル



マターンカーネル