サンプリングによるデータ駆動科学

IBIS2019 企画セッション：データ駆動科学と機械学習

福島孝治

東京大学大学総合文化研究科/先進科学研究機構
† 物質・材料研究機構
E-mail: k-hukushima@e.gcc.u-tokyo.ac.jp
URL: http://hukushimalab.u-tokyo.ac.jp/

2019/11/21

Outline

自己紹介 福島孝治 (ふくしまこうじ) 統計物理

京都生まれ

- 生まれは桃山御陵前
- 育ちは日本海側、今はなき竹野郡丹後町 (京丹後市)

1987–1991 筑波大学 第一学群 自然学類 (今はなき第一学群…)
1991–1996 筑波大学 物理学研究科 大学院生

- スピングラスの研究 (平衡統計力学，非平衡ダイナミクス)
- 拡張アンサンブル型のモンテカルロ法の提案

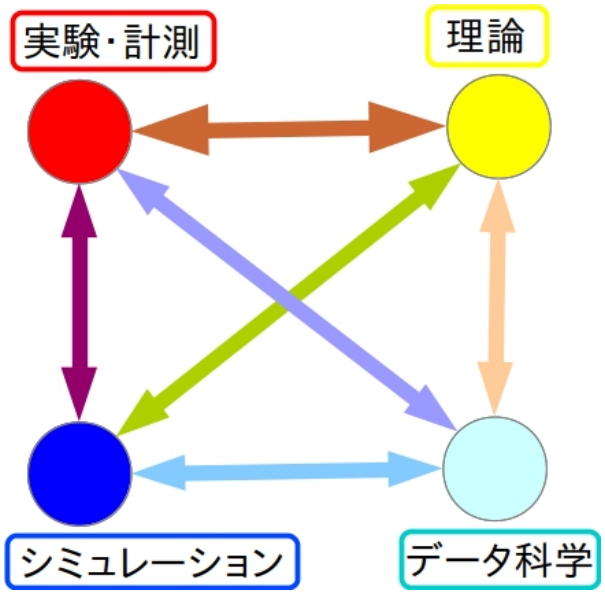
1996-2002 東京大学 物性研究所 助手 (六本木から柏へ)

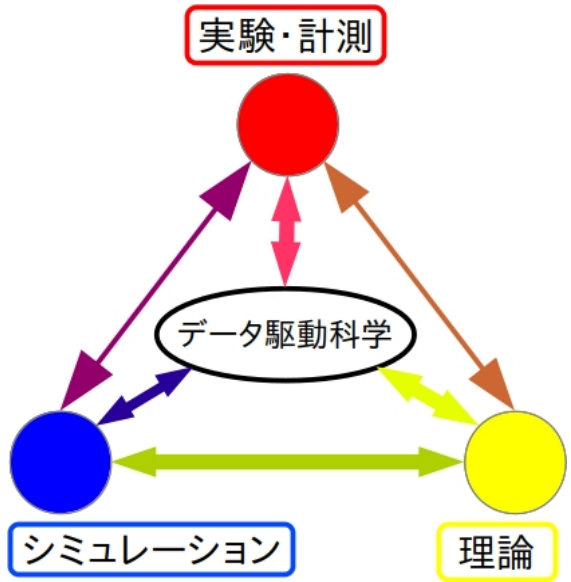
- スピングラスの相転移理論 (カイラル秩序，カオス…)
- エージング現象, 非平衡ダイナミクス，自由エネルギー計算
- 特定領域研究「確率的情報処理」SMAPIP(代表 田中和先生@東北大, 2001-2005)

2002-現在まで 東京大学 大学院総合文化研究科

- 相転移論一般
- 最適化問題の相転移など…ガラスにも興味を…
- 情報統計力学・データ駆動科学の方法論
- 特定領域研究「情報統計力学」(代表 樺島先生@東工大), 2006-2009)
- 新学術領域「スパースモデリング」(代表 岡田先生@東大), 2013-2018)
- 国立研究開発法人物質・材料研究機構 (NIMS) 招聘研究員
- 先進科学研究機構 (2019–)

データ科学と**データ駆動科学**





データ科学 Data Science

第四の科学の方法としてのデータ科学

データ駆動科学 Data-Driven Science

データの扱いを通して、自然科学の仮設検証ループを効率よく実現する
データ駆動科学は科学を**「駆動」**しうるのか?

データ科学なのかデータ駆動科学なのか...

自然科学データ

- 自然科学の実験技術は日々進歩している
  - 例えば，様々な状況でのイメージングデータ
  - 高精度化と高次元化
- 大規模数値計算もデータ生成場とも思える
  - 第一原理計算と呼ばれるできるだけ物質の根源に近い計算
  - モデルを仮定したシミュレーション (分子動力学計算)

目的：$\boldsymbol{y} \Longrightarrow \boldsymbol{x}$

- 獲得したデータ $\boldsymbol{y}$ から，その根源的な構造 $\boldsymbol{x}$ を推定したい
- 獲得したデータ $\boldsymbol{y}$ から，根源とかどうでもよいから機能 $\boldsymbol{x}$ 予測したい

ベイズ統計的アプローチ

ベイズの事後分布

$$P(\boldsymbol{x}|\boldsymbol{y}, M) = \frac{P(\boldsymbol{y}|\boldsymbol{x}, M)P(\boldsymbol{x}|M)}{Z(\boldsymbol{y}|M)}$$

- 尤度: $P(\boldsymbol{y}|\boldsymbol{x})$
  - 物理モデル+観測モデル: $\boldsymbol{x} \to \boldsymbol{y}$
  - 実験に依存して様々な物理プロセスを考慮したいところ
    - X 線回折：フーリエ変換 (線形) $\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$
    - 陽電子：シュレディンガー方程式 (非線形)$\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{\epsilon}$
- 事前分布: $P(\boldsymbol{x})$
  - 推定するパラメータに対する物理的制限や要請
- Evidence: $Z(\boldsymbol{y})$
  - 事後分布の規格化定数... これがモデルの良し悪し

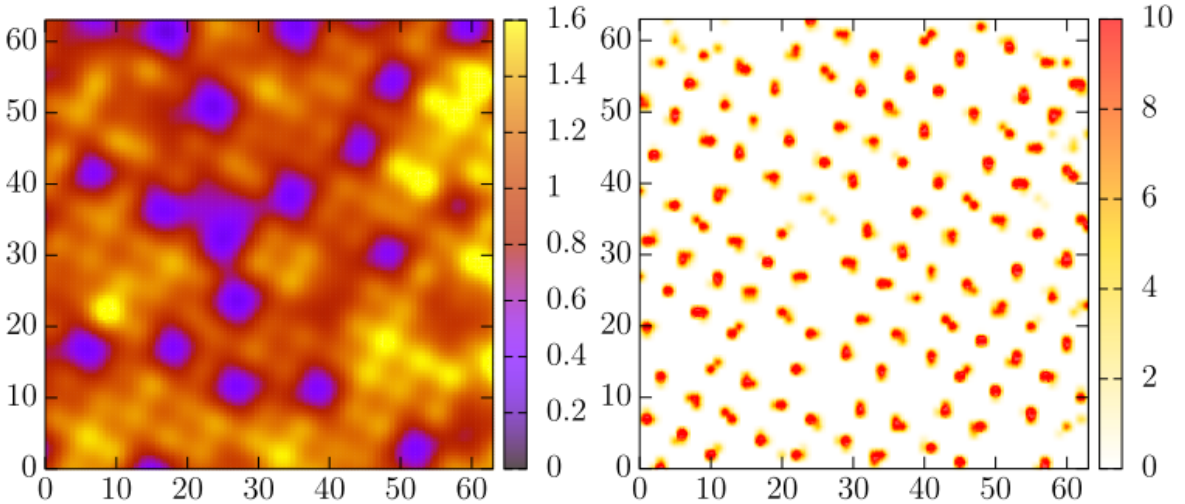
一方で... こういうこともよくわからない.

- 物理法則の知見は機械学習の中に入れるべきか否か?
- 物理法則の知見は機械学習の中に入れたいのかどうか?

Case 1: Scanning Tunneling Microscopy topography data analysis

M.J.Miyama and KH (2018)

- data $\boldsymbol{y}$: real-space topography data of $SrVO_3$
- parameter $\boldsymbol{x}$: atom position in real space
- modeling : $\boldsymbol{y} = \hat{A}\boldsymbol{x} + \boldsymbol{\epsilon}$ with point-spread-function A_{ij} and noise $\boldsymbol{\epsilon}$
- sparse modeling for $\boldsymbol{x}$



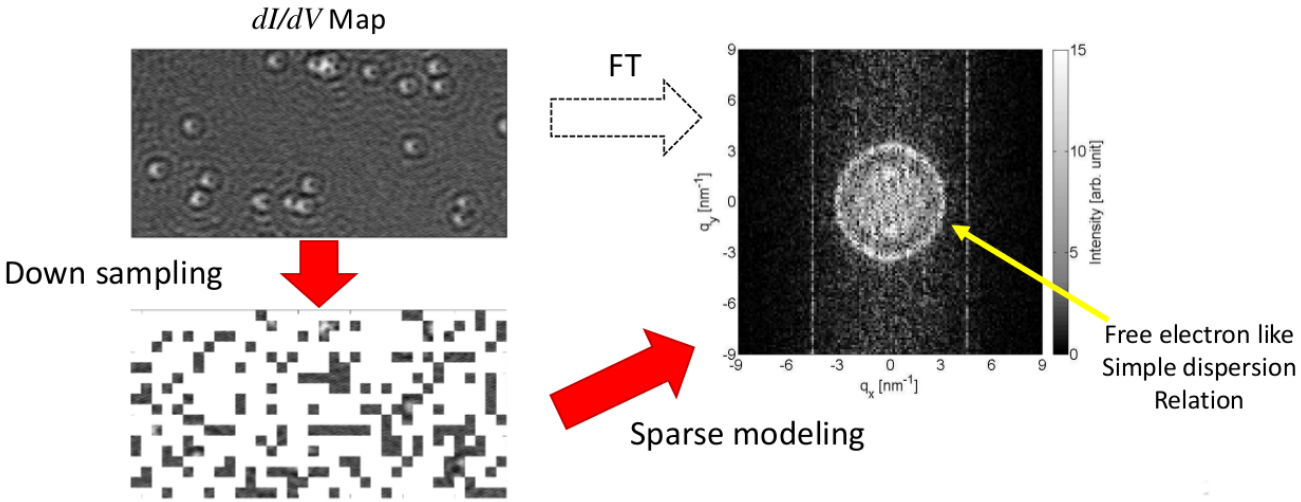
(a) Original data y_i (b) Estimated amplitude x_i

Vacancy rate and lattice distortion by vacancies are quantitatively estimated.

Case 2: Scanning Tunneling Spectroscopy

Y.Nakanishi-Ohno et al (2016)

- data $\boldsymbol{y}$: real-space imaging data of current-voltage relation
- parameter $\boldsymbol{x}$: dispersion relation in Fourier space
- modeling : $\boldsymbol{y} = \hat{A}\boldsymbol{x} + \boldsymbol{\epsilon}$ with Fourier transformation matrix $\hat{A}$.
- sparse modeling for $\boldsymbol{x}$

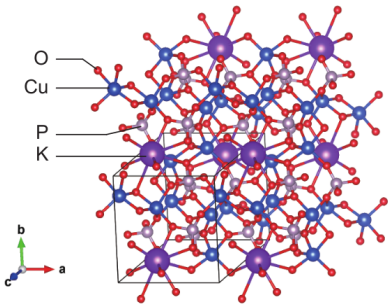


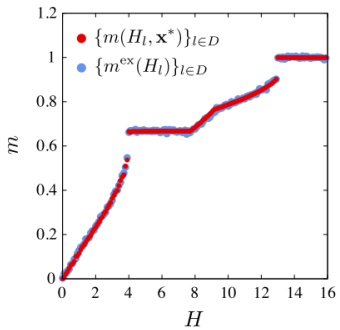
the same level of reconstruction is possible from the down-sampling data.

Case 3: Hamiltonian reconstruction from experimental data

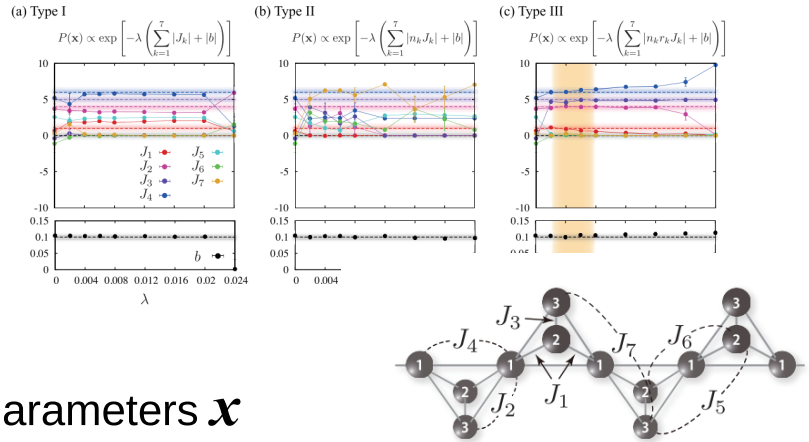
R.Tamura and KH (2017, 2018, 2019...)

- data $\boldsymbol{y}$: experimentally observed physical quantities
- parameter $\boldsymbol{x}$: parameters in model Hamiltonian
- modeling : $\boldsymbol{y}$ is calculated by statistical-mechanical many-body calculation with a non-linear function $g(\boldsymbol{x})$ as $\boldsymbol{y} = g(\boldsymbol{x}) + \boldsymbol{\epsilon}$.
- Monte Carlo sampling from the posterior probability $P(\boldsymbol{x}|\boldsymbol{y})$.





Different type of Measurements $\boldsymbol{y}$



Parameters $\boldsymbol{x}$ in model Hamiltonian

今日，議論したい問題 I

ノイズ付き線形方程式

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

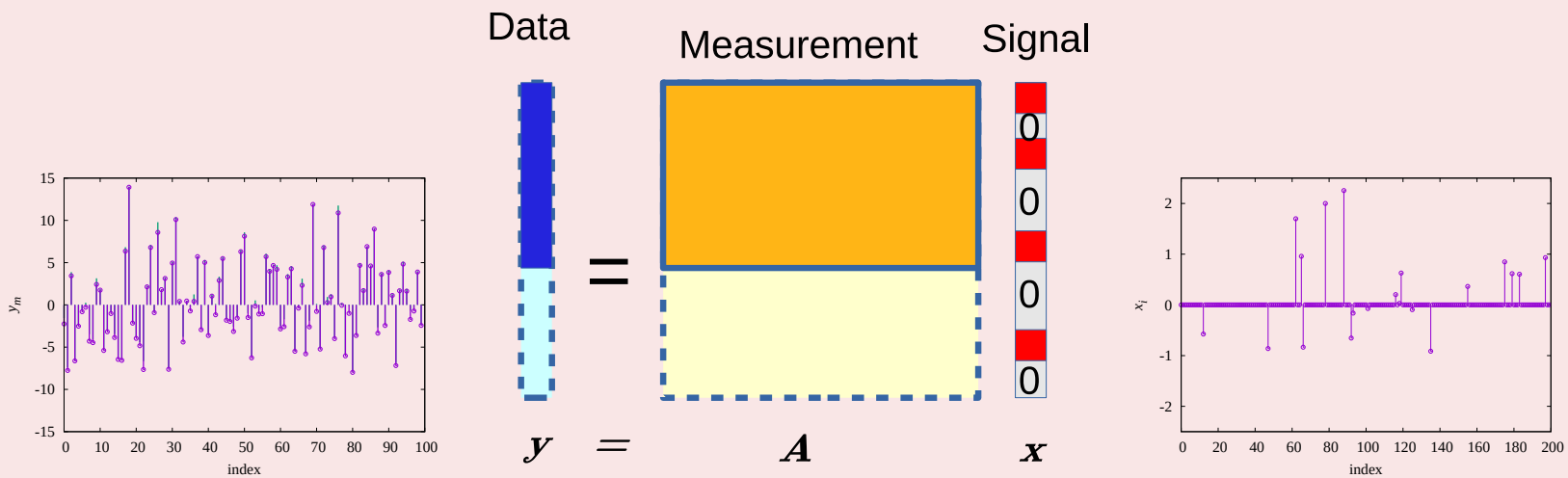
- $\boldsymbol{y}$: 観測データ (M 次元ベクトル)
- A: 観測行列 ($M \times N$)
  - 計測系によって与えられる．e.g. フーリエ行列
- $\boldsymbol{x}$: 信号源データ (N 次元ベクトル)
- $\boldsymbol{\epsilon}$: ノイズデータ (unknown N 次元ベクトル)

問い

1. 観測データ $\boldsymbol{y}$ から信号源 $\boldsymbol{x}$ を推定せよ
   - 一般に，$M < N$ なら，不定問題
   - この範疇に属する多くの実験系：MRI,NMR,STS,$\cdots$

今日，議論したい問題 II

Sparse modeling



問い *

1. 観測データ $\boldsymbol{y}$ から信号源 $\boldsymbol{x}$ を推定せよ．ただし，$\boldsymbol{x}$ の非ゼロ要素 $N_{\boldsymbol{x}}$ は M より十分少ないと仮定してよい．
2. 非ゼロ要素数 $N_{\boldsymbol{x}}$ も評価せよ

今日，議論したい問題 III

- 事後分布

$$P(\boldsymbol{x}|\boldsymbol{y}, N_{\boldsymbol{x}}) = \frac{P(\boldsymbol{y}|\boldsymbol{x}, N_{\boldsymbol{x}})P(\boldsymbol{x}|N_{\boldsymbol{x}})}{Z(\boldsymbol{y}|N_{\boldsymbol{x}})}$$

- $P(\boldsymbol{y}|\boldsymbol{x}, N_{\boldsymbol{x}})$: ガウスノイズを仮定した尤度

$$P(\boldsymbol{y}|\boldsymbol{x}, N_{\boldsymbol{x}}) = \left(\frac{\beta}{2\pi}\right)^{M/2} \exp\left(-\frac{\beta}{2}\left(\boldsymbol{y} - A\boldsymbol{x}\right)^2\right)$$

  - β: 逆温度 (ノイズの強さ)
  - $\boldsymbol{x} = \boldsymbol{z} \circ \boldsymbol{c}, \quad z_n \in \mathcal{R}, \quad c_n \in \{0, 1\}$
  - $\boldsymbol{c}$ の割当問題としての事後分布最大化が NP hard だからって，ビビってる場合じゃないよね.
- 事前分布: スパース性をハードに.
- $Z(\boldsymbol{y}|N_{\boldsymbol{x}})$: 規格化定数 (エビデンス)

事後分布からのサンプリングへ

解法に求められる要件

1. 将来的には，**線形問題から非線形問題，単峰から多峰分布へ．対応.**
   - 勾配ベースの方法では局所解のみしかわからない$\Longrightarrow y = g(x, \Theta) + \epsilon$
   - 局所解ではなくて，大域解をちゃんと求めたい
2. **規格化定数が計算できる** $Z(y)$
   - 分布の最大値だけを求めるなら最適化手法でよい.
3. **大規模データへの対応のための大規模計算**
   - large N 問題
     - 変数 (x) の次元：次元の呪い
     - データ (y) の次元: 1 ステップの計算量
   - 時代は並列計算機

モンテカルロ・サンプリング法

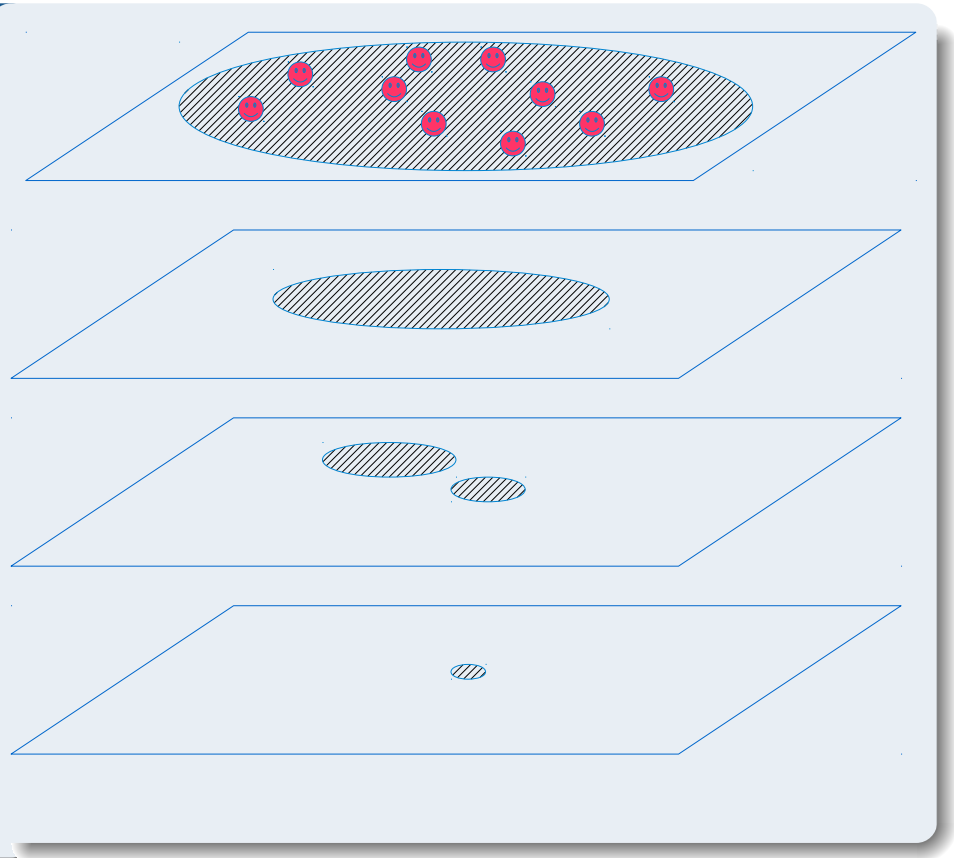
- 高次元確率分布からの汎用サンプリング法
- 規格化定数の計算
- 並列計算の並列度数...$10^5 \sim 10^6$ $\Longleftarrow$ **ここがポイント**

Population Annealing

Hukushima-Iba (2003)

ステップ 0: 初期条件

- 十分高温 $1/\beta_0$ から.
- M 個のレプリカ $x_0^k, (k = 1, 2, \cdots, M)$ をランダム生成
- レプリカ k の重み W_k を初期化: $W_k^{(0)} = 1$.



ステップ 1: MC 更新とアニール

- $\beta = \beta_i$ で，それぞれのレプリカの状態更新 $\{x_i^k\} \Longrightarrow \{x_{i+1}^k\}$
- アニール : $\beta_i \Longrightarrow \beta_{i+1} = \beta_i + \Delta\beta$

ステップ 2: 重みの更新 $\{W_k^{(i)}\} \Longrightarrow \{W_k^{(i+1)}\}$

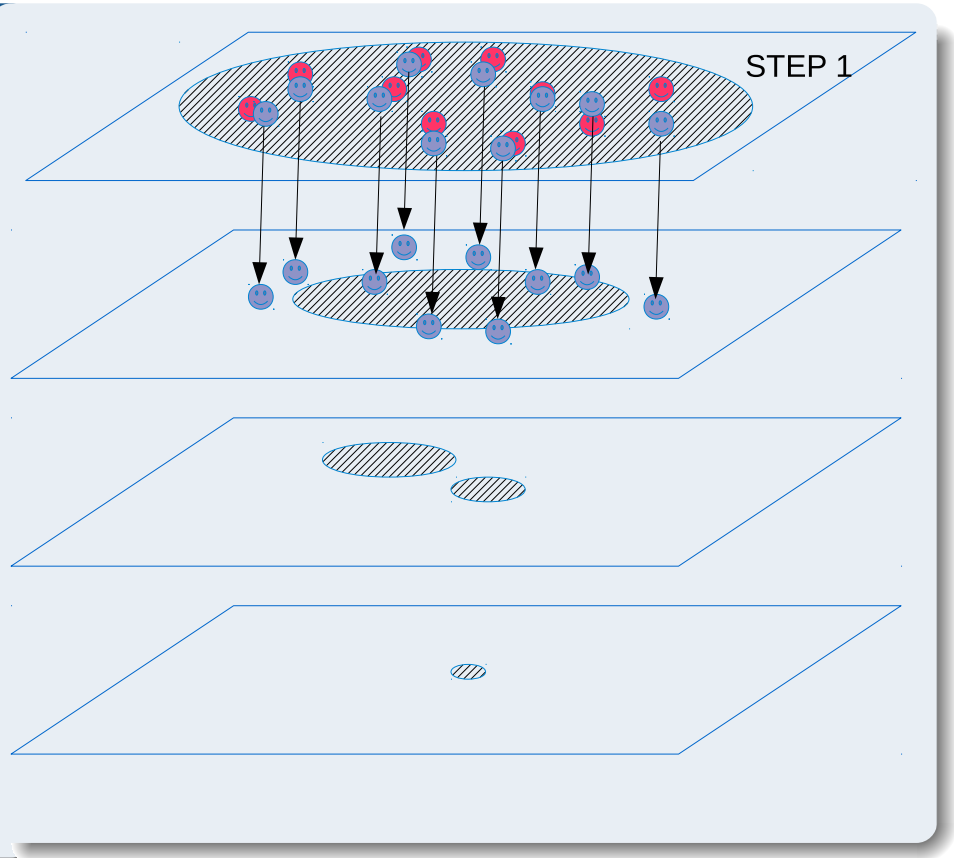
- $W_k^{(i+1)} = \dfrac{P_{\beta_{i+1}}(x_{i+1}^k)}{P_{\beta_i}(x_{i+1}^k)} W_k^{(i)}(x_0^k, x_1^k, \cdots, x_i^k)$

Population Annealing

Hukushima-Iba (2003)

ステップ 0: 初期条件

- 十分高温 $1/\beta_0$ から.
- M 個のレプリカ $x_0^k, (k = 1, 2, \cdots, M)$ をランダム生成
- レプリカ k の重み W_k を初期化: $W_k^{(0)} = 1$.



ステップ 1: MC 更新とアニール

- $\beta = \beta_i$ で，それぞれのレプリカの状態更新 $\{x_i^k\} \Longrightarrow \{x_{i+1}^k\}$
- アニール : $\beta_i \Longrightarrow \beta_{i+1} = \beta_i + \Delta\beta$

ステップ 2: 重みの更新 $\{W_k^{(i)}\} \Longrightarrow \{W_k^{(i+1)}\}$

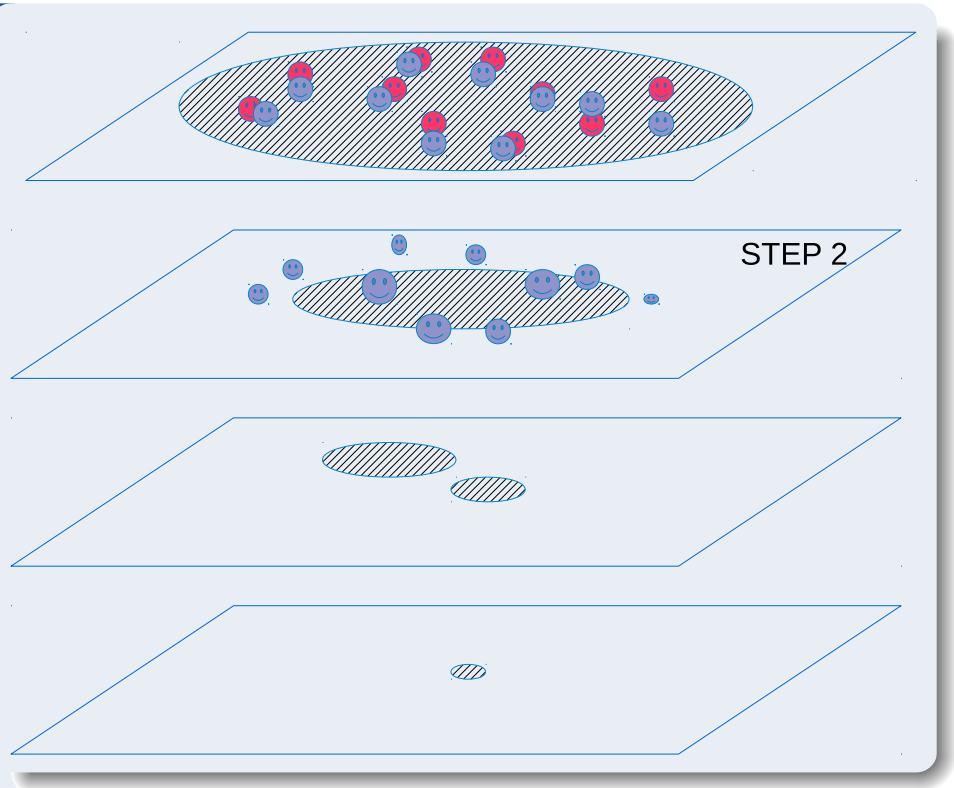
- $W_k^{(i+1)} = \dfrac{P_{\beta_{i+1}}(x_{i+1}^k)}{P_{\beta_i}(x_{i+1}^k)} W_k^{(i)}(x_0^k, x_1^k, \cdots, x_i^k)$

Population Annealing

Hukushima-Iba (2003)

ステップ 0: 初期条件

- 十分高温 $1/\beta_0$ から.
- M 個のレプリカ $x_0^k, (k = 1, 2, \cdots, M)$ をランダム生成
- レプリカ k の重み W_k を初期化: $W_k^{(0)} = 1$.



ステップ 1: MC 更新とアニール

- $\beta = \beta_i$ で，それぞれのレプリカの状態更新 $\{x_i^k\} \Longrightarrow \{x_{i+1}^k\}$
- アニール : $\beta_i \Longrightarrow \beta_{i+1} = \beta_i + \Delta\beta$

ステップ 2: 重みの更新 $\{W_k^{(i)}\} \Longrightarrow \{W_k^{(i+1)}\}$

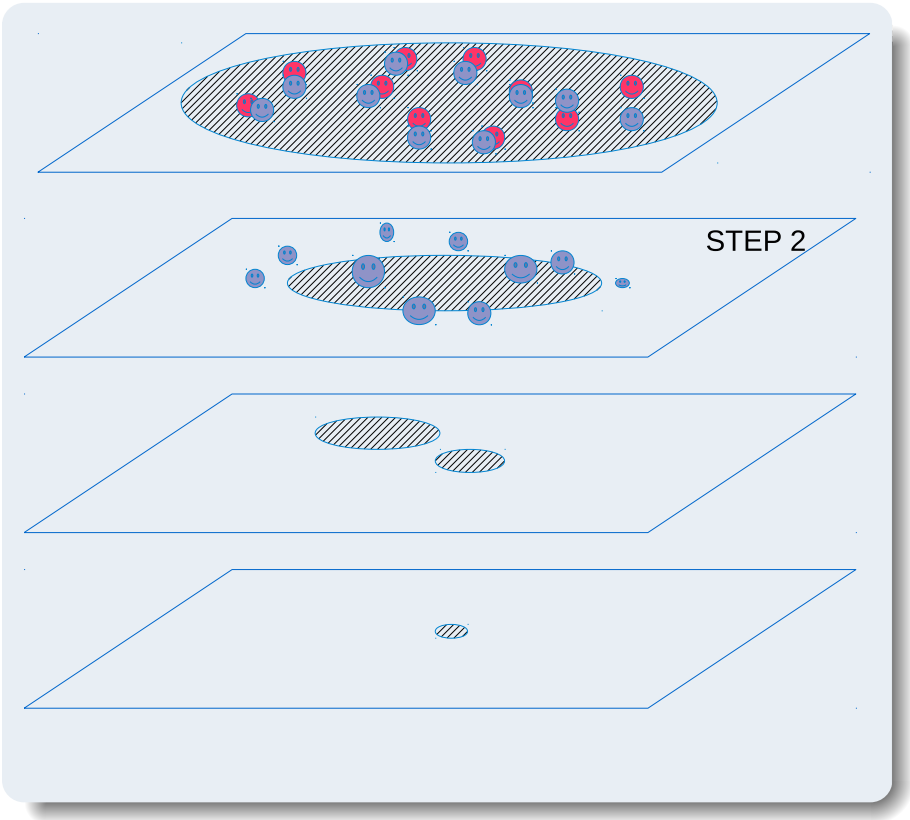
- $W_k^{(i+1)} = \dfrac{P_{\beta_{i+1}}(x_{i+1}^k)}{P_{\beta_i}(x_{i+1}^k)} W_k^{(i)}(x_0^k, x_1^k, \cdots, x_i^k)$

Population Annealing ステップ３：リサンプリング

1. 各レプリカの重みから確率

$$P^k = \frac{W_k^{(i)}}{\sum_k W_k^{(i)}}$$

2. その確率に従って，レプリカ x_i^k をリサンプリング (復元抽出)
   - 重みが小さいレプリカは消される
   - 重みが大きいレプリカは分裂
3. 重みはリセット: $W_k^{(i)} = 1$.



期待値 $\langle A \rangle$ at $1/\beta_j$ と規格化定数 Z

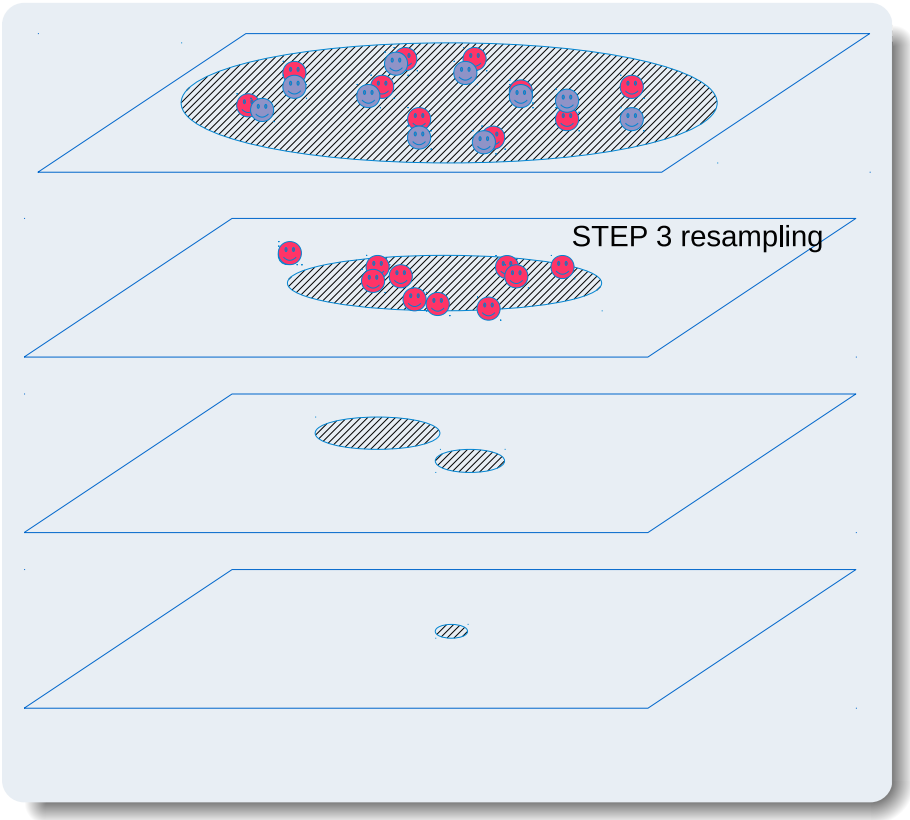
$$\langle A \rangle_{\beta_j} = \frac{\sum_{k=1}^M A(x_j^k) W_k^{(j-1)}}{\sum_{k=1}^M W_k^{(j-1)}} = \frac{\langle A W^{(j-1)} \rangle_{\text{path}}}{\langle W^{(j-1)} \rangle_{\text{path}}}, \quad \langle W^{j-1} \rangle_{\text{path}} = \frac{Z_j}{Z_0}$$

Population Annealing ステップ 3 : リサンプリング

1. 各レプリカの重みから確率

$$P^k = \frac{W_k^{(i)}}{\sum_k W_k^{(i)}}$$

2. その確率に従って，レプリカ x_i^k をリサンプリング (復元抽出)
   - 重みが小さいレプリカは消される
   - 重みが大きいレプリカは分裂
3. 重みはリセット: $W_k^{(i)} = 1$.



期待値 $\langle A \rangle$ at $1/\beta_j$ と規格化定数 Z

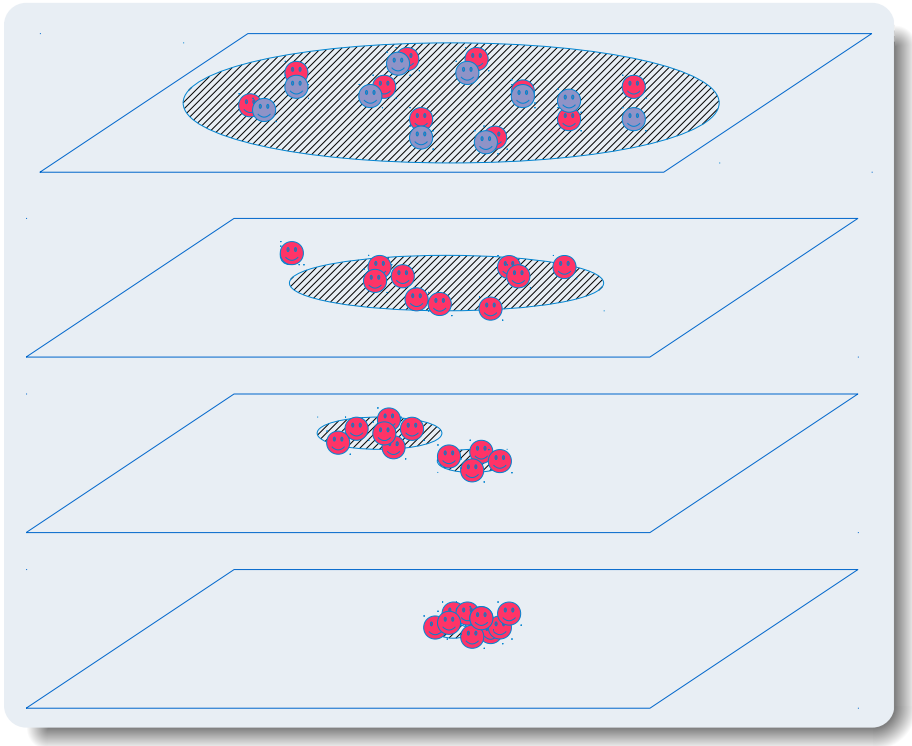
$$\langle A \rangle_{\beta_j} = \frac{\sum_{k=1}^M A(x_j^k) W_k^{(j-1)}}{\sum_{k=1}^M W_k^{(j-1)}} = \frac{\langle A W^{(j-1)} \rangle_{\text{path}}}{\langle W^{(j-1)} \rangle_{\text{path}}}, \quad \langle W^{j-1} \rangle_{\text{path}} = \frac{Z_j}{Z_0}$$

Population Annealing ステップ３：リサンプリング

1. 各レプリカの重みから確率

$$P^k = \frac{W_k^{(i)}}{\sum_k W_k^{(i)}}$$

2. その確率に従って，レプリカ x_i^k をリサンプリング (復元抽出)
   - 重みが小さいレプリカは消される
   - 重みが大きいレプリカは分裂
3. 重みはリセット: $W_k^{(i)} = 1$.



期待値 $\langle A \rangle$ at $1/\beta_j$ と規格化定数 Z

$$\langle A \rangle_{\beta_j} = \frac{\sum_{k=1}^M A(x_j^k) W_k^{(j-1)}}{\sum_{k=1}^M W_k^{(j-1)}} = \frac{\langle A W^{(j-1)} \rangle_{\text{path}}}{\langle W^{(j-1)} \rangle_{\text{path}}}, \quad \langle W^{j-1} \rangle_{\text{path}} = \frac{Z_j}{Z_0}$$

Population Annealing について

利点

- 規格化定数の計算がほぼただ
- 並列が容易?．特に，超並列向き
- 重みの再重率化で，LOOCV も計算可能
- サンプルする確率分布に対する柔軟性
  - 拡張アンサンブル法でできることは PA でもだいたいできる．
  - 状態密度は交換法系より一手間軽い

不利点

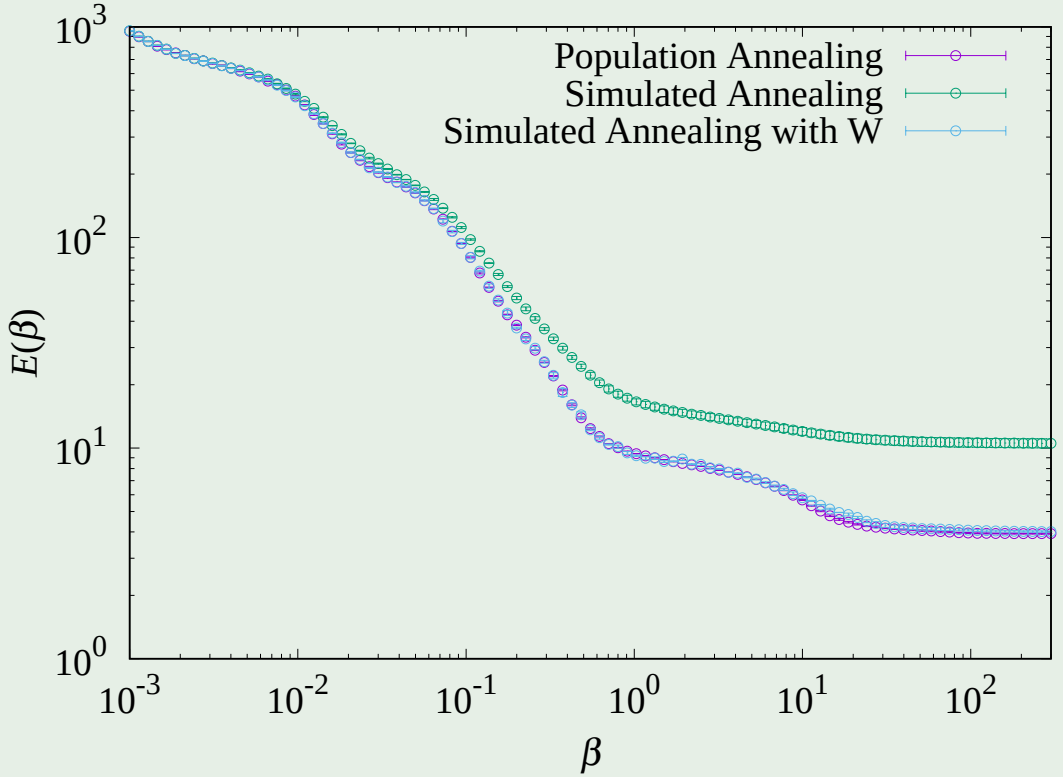
- マルコフ連鎖 MC と比較して本当によいか?は不明
- そもそも理論的背景は MCMC より貧弱

数値計算結果

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

- $N = 200,\ M = 100$
- $N_x = 20$
- A: iid Gaussian
- noise
- 粒子数: 1000

Energy vs β



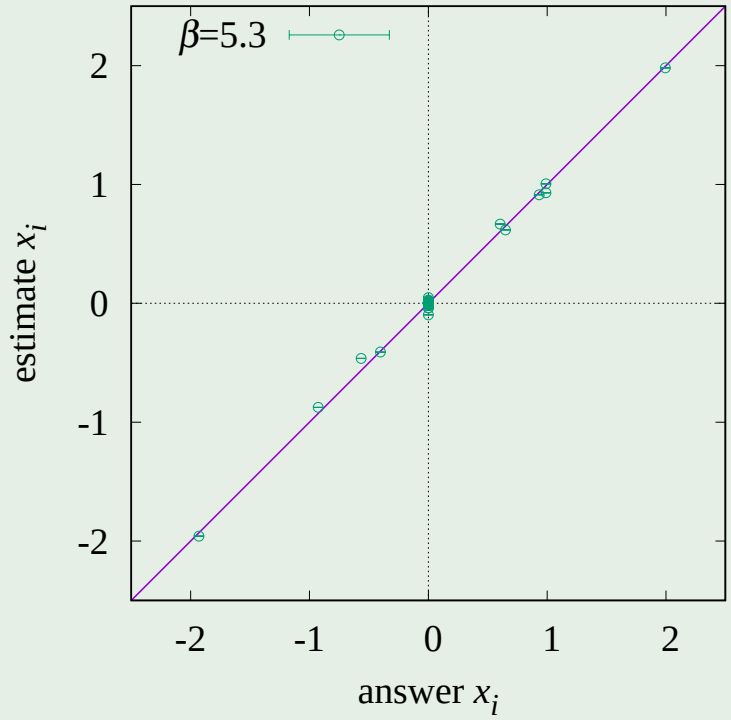
SA は失敗する．重みをつけるとまだ耐える．

数値計算結果

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

- $N = 200,\ M = 100$
- $N_x = 20$
- A: iid Gaussian
- noise
- 粒子数: 1000

estimate of x_i



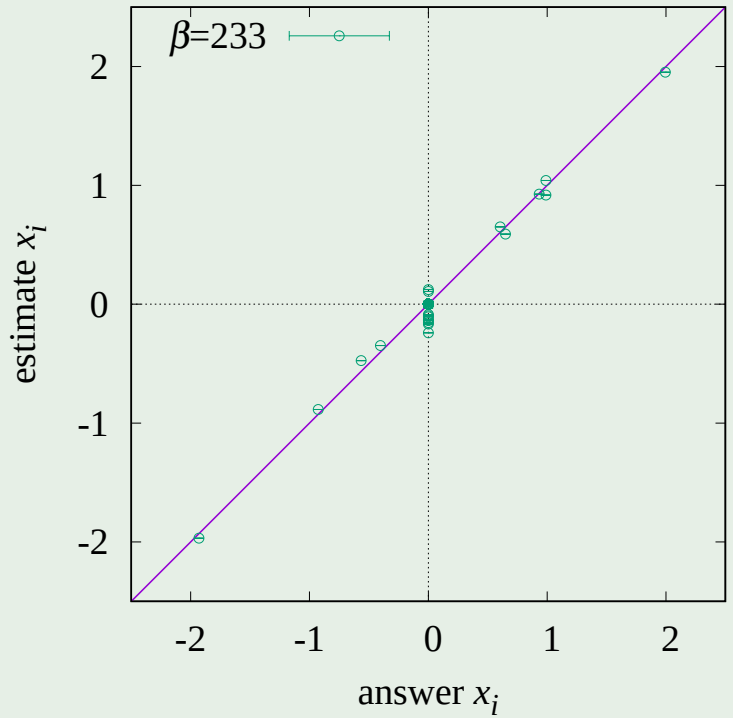
ノイズ (β) の大きさがほどほどだとうまく推定できている

数値計算結果

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

- $N = 200,\ M = 100$
- $N_x = 20$
- A: iid Gaussian
- noise
- 粒子数: 1000

estimate of x_i



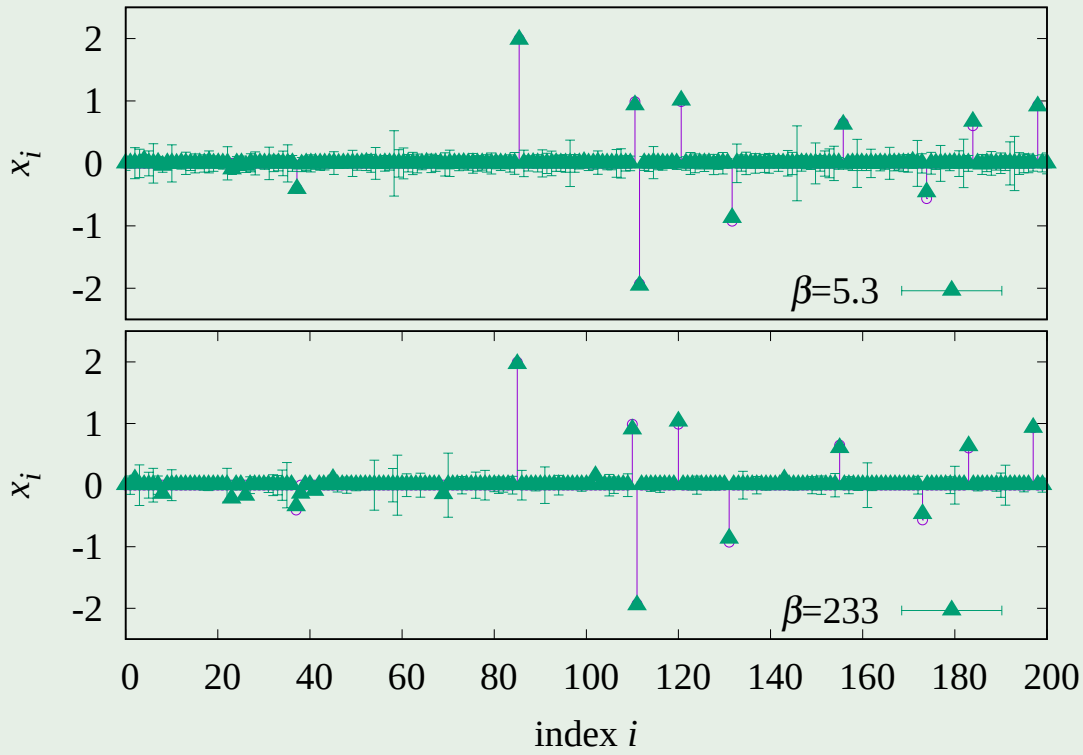
ノイズを過小評価?!

数値計算結果

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

- $N = 200,\ M = 100$
- $N_x = 20$
- A: iid Gaussian
- noise
- 粒子数: 1000

estimate of x_i

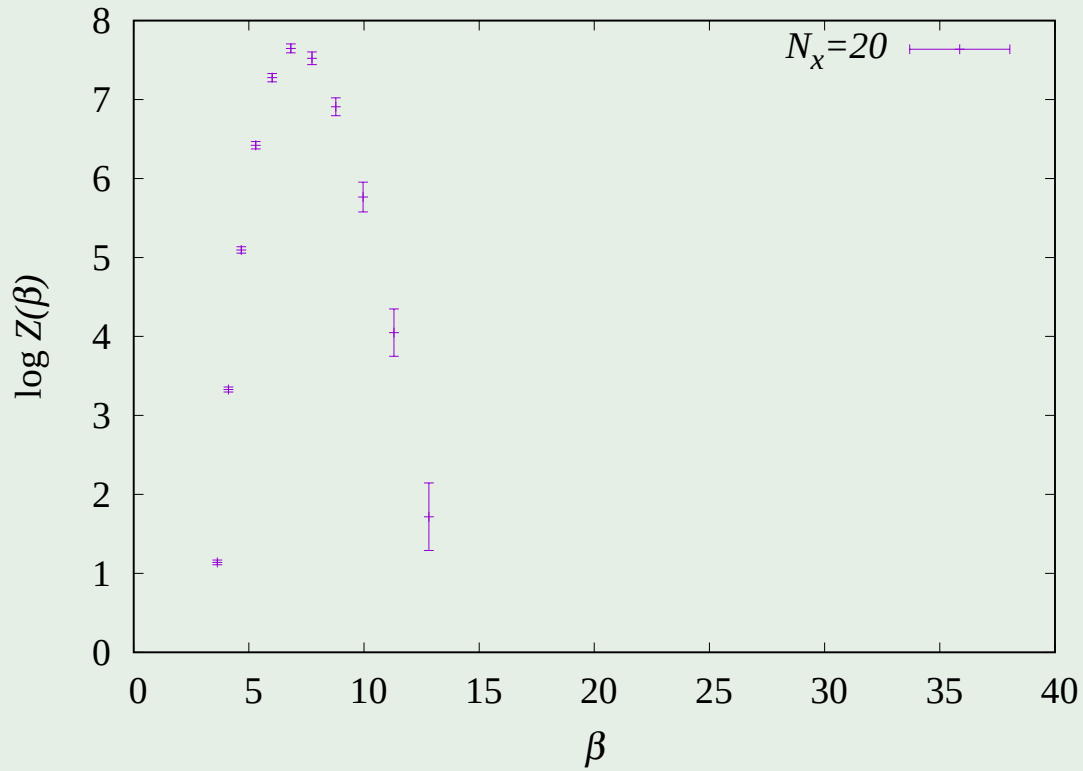


数値計算結果

$$\boldsymbol{y} = A\boldsymbol{x} + \boldsymbol{\epsilon}$$

- $N = 200,\ M = 100$
- $N_x = 20$
- A: iid Gaussian
- noise
- 粒子数: 1000

Evidence

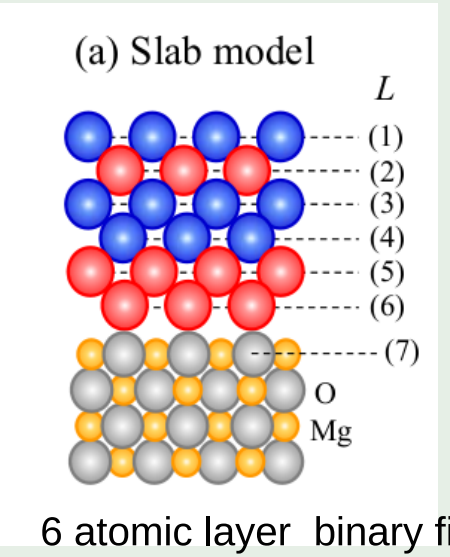


ここからノイズの大きさも推定できる.

第一原理電子状態計算の解析

物質科学の分野での第一原理計算は，物質の電子状態を知る理論的方法として広く使われている．

磁性薄膜 (AuFe) on MgO: Au と Fe の配置で物性は大きく異なる

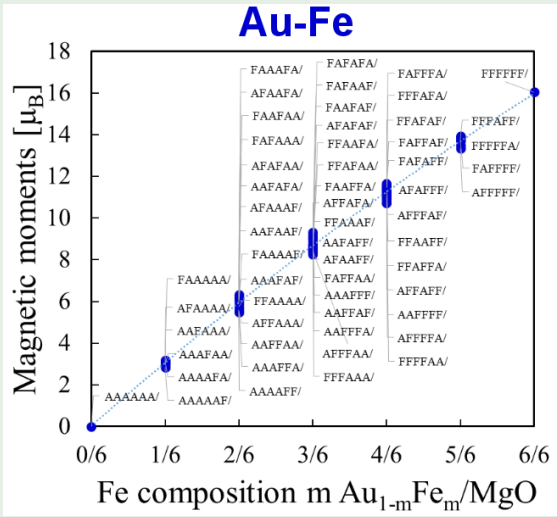


6 atomic layer binary film

AAAAAA/MgO
AAAAAF/MgO
AAAAFA/MgO
AAAAFF/MgO
AAAFAA/MgO

FFFAAF/MgO
FFFAFF/MgO
FFFFAA/MgO
FFFFAF/MgO
FFFFFA/MgO
FFFFFF/MgO

First-principle calculation

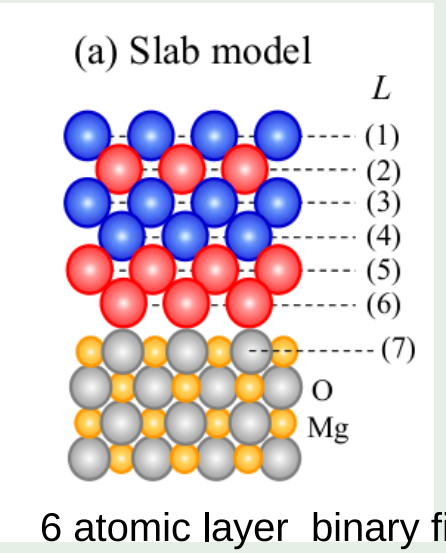


磁気モーメントは鉄原子層の個数に比例する simple.

第一原理電子状態計算の解析

物質科学の分野での第一原理計算は，物質の電子状態を知る理論的方法として広く使われている．

磁性薄膜 (AuFe) on MgO: Au と Fe の配置で物性は大きく異なる

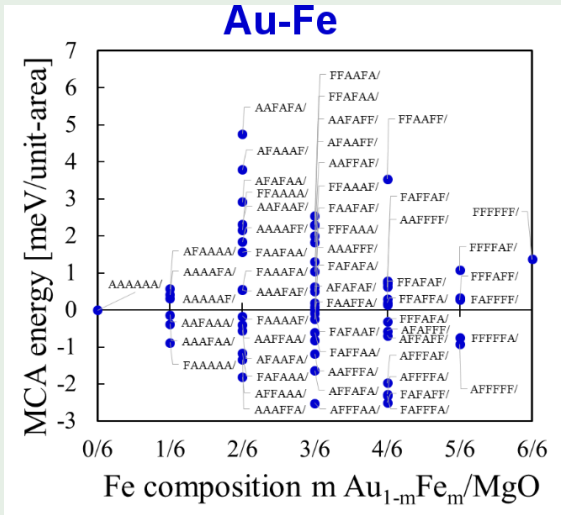


6 atomic layer binary film

AAAAAA/MgO
AAAAAF/MgO
AAAAFA/MgO
AAAAFF/MgO
AAAFAA/MgO

FFFAAF/MgO
FFFAFF/MgO
FFFFAA/MgO
FFFFAF/MgO
FFFFFA/MgO
FFFFFF/MgO

First-principle calculation



結晶磁気異方性 (MCA) は激しく依存する．

我々の目的は...

MCA を増大させる物理的な原理を理解したい...
高機能スピンデバイスの設計のために重要

この解析での記述子

原子層の表現

- 原子層を表す変数：$\sigma_i = +1 \quad (\mathrm{Au}), \quad \mathrm{or} -1 \quad (\mathrm{Fe}).$
- 原子層：$\boldsymbol{\sigma} = (\sigma_1, \sigma_2, \cdots, \sigma_6)$

$$\text{e.g., AuAuFeAuFeFe} \Longrightarrow (+1, +1, -1, +1, -1, -1)$$

- 相関関数表示 $\boldsymbol{\xi} = (\xi_1, \xi_2, \cdots, \xi_{64})$
  - 一体相関 $\xi_1 = \langle \sigma_1 \rangle$, ...
  - 二体相関 $\xi_7 = \langle \sigma_1 \sigma_2 \rangle$, ...
  - $\cdots$, up to 六体相関 $\xi_{64} = \langle \sigma_1 \sigma_2 \cdots \sigma_6 \rangle$
- 線形回帰: 第一原理計算で得られた MCA データ $\boldsymbol{y}$ から，説明変数の係数 $\boldsymbol{x}$ を推定

$$y(\boldsymbol{\xi}) = x_1 \xi_1 + x_2 \xi_2 + \cdots + x_{64} \xi_{64}$$

どの相関変数が MCA の増大に重要か?

もし $\xi_8 = \langle \sigma_2 \sigma_3 \rangle$ が正の係数をもって必要となると，... (*AuAu***) or (*FeFe***) が重要とわかる.

解析結果

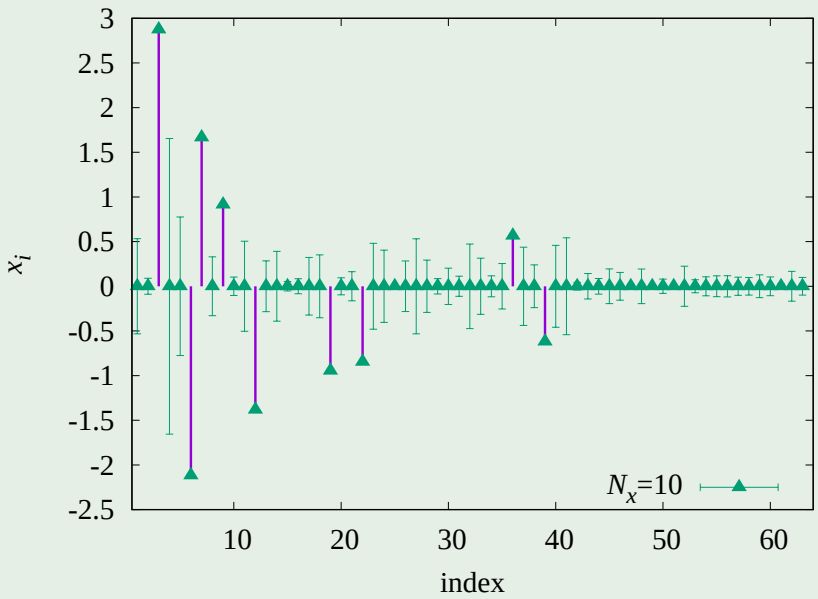
非ゼロ要素 N_x を止めて，全数検索を行うことは実質的に困難：
ここではポピュレーションアニーリングによるモンテカルロ・サンプリングで計算

relevant var.

- positive ξ_3 (**Au***)
- negative ξ_6 (*****Fe)
- positive ξ_7 ((??)****)
- positive ξ_9 (?**?**)
- negative ξ_{12} (*??***)
- ...

A consistent alignment:
FeFeAuAuFeFe

estimate of x_i



ここからわかるメッセージ:
境界層には Fe 原子層を置くのがよい... ようだ.

まとめ

- データ駆動科学にとって不確定度の評価のためにサンプリングが必要
- 一つのサンプリング法として Population annealing (PA)
  - ポピュレーション型モンテカルロ法
  - ◯ それぞれの粒子は**ほぼ**独立な計算なので，高並列度への対応
  - △　「ほぼ」は resampling の部分．ここはやや面倒くさい
  - 規格化定数は追加計算なしに評価
  - マルコフ連鎖ほどの理論がないのが現状.
- 次のデータを取得するという意味でのサンプリングも重要．ベイズ最適化など.

謝辞

- 共同研究：長野くん@東大駒場
- 共同研究：中西さん@東大駒場，岡田さん@東大，田村さん@NIMS, 木野さん@ NIMS，中村さん@三重大, 観山さん@東北大
- 議論：J.Machta, M. Weigel, W. Wang