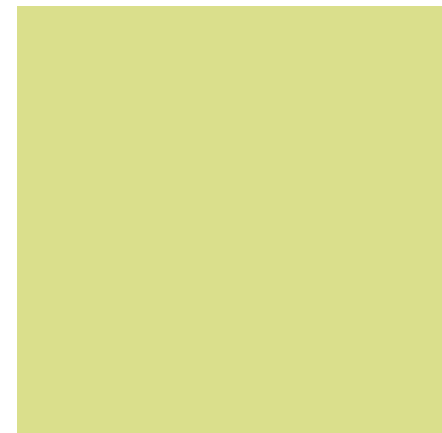




大阪市立大学電子・物理工学特別講義

今日からできる スパースモデリング

京都大学大学院情報学研究科システム科学専攻

大関 真之

QUANTUM
ANNEALING

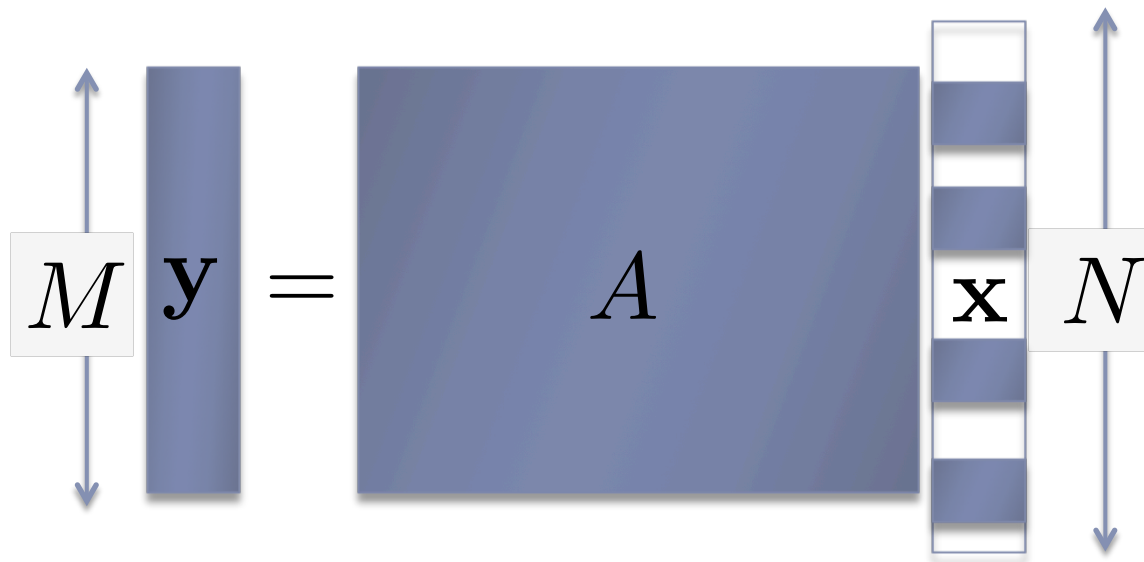


MACHINE
LEARNING

Sparse Modeling

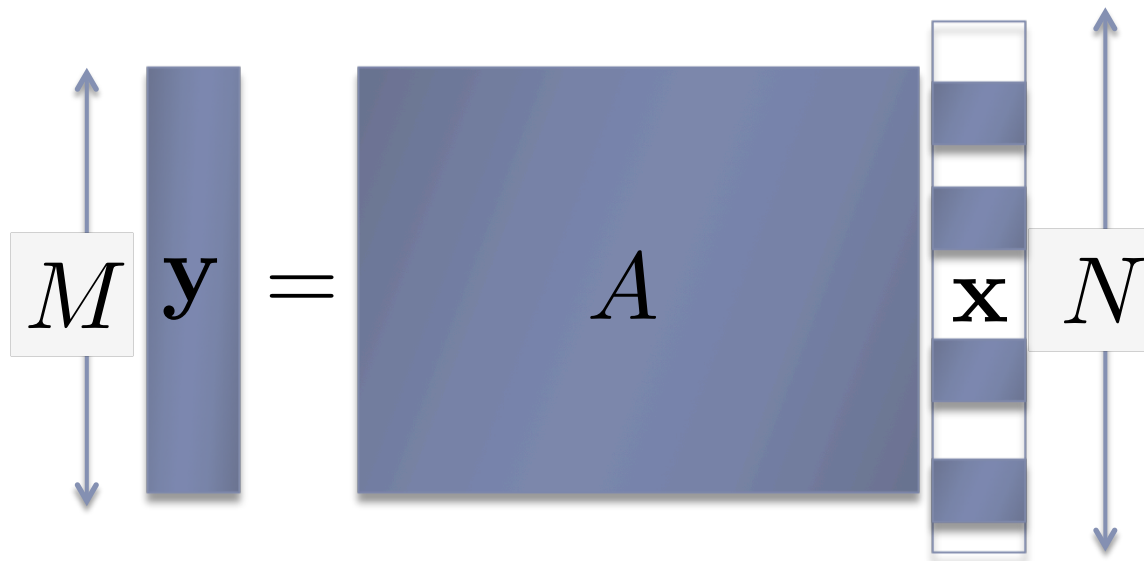


京都大学
KYOTO UNIVERSITY



$$y = Ax$$

を満たす解を求めよ



$$y = Ax$$

を満たす解を求めよ
できる！

スパース解推定

- ▶ L0ノルム最小化によるスパース解推定
 - ▶ 以下の最小化問題でスパース解を探索しよう！

$$\min_{\mathbf{x}} \|\mathbf{x}\|_0 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}$$

- ▶ 非零成分の個数が小さい（スパース）で方程式を満たすものを探す
 - ▶ 残念ながら計算量的に困難
- ▶ **L1**ノルム最小化によるスパース解推定
 - ▶ 代わりに次の最小化問題でスパース解を探索しよう！

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}$$

- ▶ 零が多ければ小さくなる + **大きさも小さくなりがち**
 - ▶ 計算量は非常に軽い（Nの3乗程度）



▶ L_1 ノルム最小化によるスパース解推定法

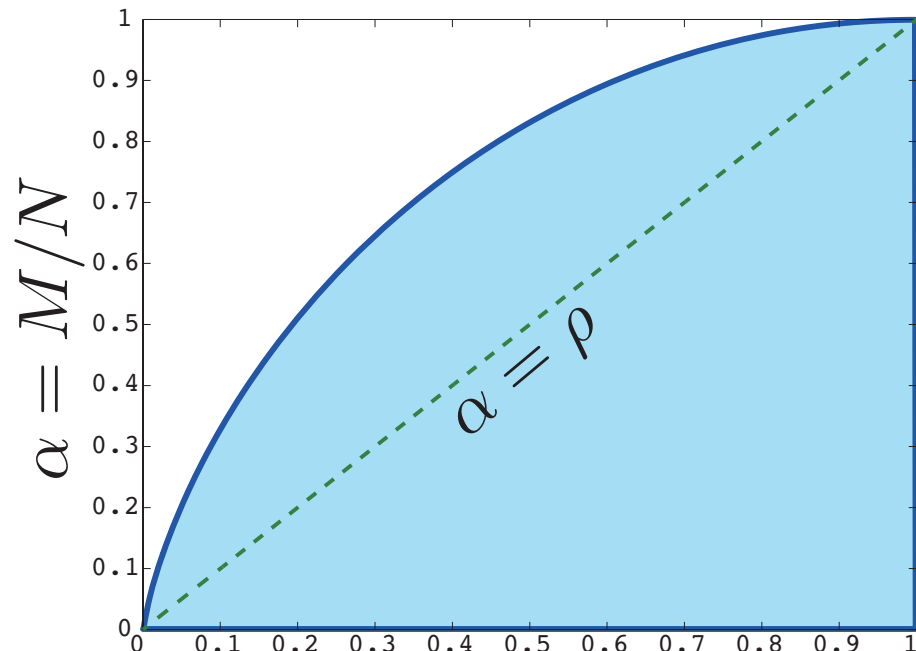
▶ 基底追跡 (Basis Pursuit)

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}$$

$$\frac{1}{\alpha} = 1 + \sqrt{\frac{\pi}{2}} t e^{\frac{t^2}{2}} \{1 - 2Q(t)\}$$

$$\frac{\rho}{1 - \rho} = 2 \left(\frac{e^{\frac{t^2}{2}}}{t\sqrt{2\pi}} - Q(t) \right)$$

$$Q(t) = \int_t^\infty \frac{e^{-\frac{x^2}{2}}}{\sqrt{2\pi}} dx$$



$\rho = K/N$
Sparse Modeling

ノイズ有り圧縮センシング

- ▶ 観測結果にはノイズがつきもの？！

- ▶ 加法的ノイズが混入した場合

$$\mathbf{y} = A\mathbf{x} + \sigma_0 \mathbf{w}$$

- ▶ 等式制約はもはや成立しない

- ▶ LASSO (Least Absolute Shrinkage and Selection Operators)

$$\min_{\mathbf{x}} \left\{ \|\mathbf{y} - A\mathbf{x}\|_2^2 \right\} \quad \text{s.t.} \quad \|\mathbf{x}\|_1 \leq a$$

- ▶ ラグランジュ未定乗数法により等価な最適化問題

$$\min_{\mathbf{x}} \left\{ \frac{1}{2\lambda} \|\mathbf{y} - A\mathbf{x}\|_2^2 + \|\mathbf{x}\|_1 \right\}$$



LASSOにメジャライザー最小化

- ▶ LASSOの場合 $g(\mathbf{x}) = \frac{1}{2\lambda} \|\mathbf{y} - A\mathbf{x}\|_2^2$

$$q_L(\mathbf{x}, \mathbf{v}) = g(\mathbf{v}) + \frac{A^T A}{\lambda} (\mathbf{y} - A\mathbf{v}) (\mathbf{x} - \mathbf{v}) + \frac{L}{2} \|\mathbf{x} - \mathbf{v}\|_2^2$$

- ▶ 平方完成してメジャライザー+L1ノルム最小解を求める

$$\mathbf{x}[t+1] = \arg \min_{\mathbf{x}} \left\{ \frac{L}{2} \left\{ \mathbf{x} - \left(\mathbf{x}[t] + \frac{1}{L\lambda} A^T (\mathbf{y} - A\mathbf{x}[t]) \right) \right\}^2 + \|\mathbf{x}\|_1 \right\}$$

- ▶ 分離性を持つ形であるので軟判定しきい値関数を適用するだけ.

$$\mathbf{x}[t+1] = S_{1/L} \left(\mathbf{x}[t] + \frac{1}{L\lambda} A^T (\mathbf{y} - A\mathbf{x}[t]) \right)$$

- ▶ ノイズなし圧縮センシング
 - ▶ 基底追跡

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}_0$$

- ▶ 制約条件ありの最適化問題
 - ▶ ラグランジュ未定乗数

$$\min_{\mathbf{x}} \left\{ \|\mathbf{x}\|_1 + (\mathbf{h}[t])^T (\mathbf{y} - A\mathbf{x}) \right\}$$

- 最小化のときは正の乗数、最大化のときは負の乗数
- 未定乗数の更新と最適化にかかる変数の更新でやや不安定

- ▶ 罰金法

$$\min_{\mathbf{x}} \left\{ \|\mathbf{x}\|_1 + \frac{\mu}{2} \|\mathbf{y} - A\mathbf{x}\|_2^2 \right\}$$

- 罰金係数を徐々に大きくしていかなければならない



統計力学的アプローチの出発点

- ▶ ミクロな世界とマクロな世界を結ぶ処方箋
 - ▶ ハミルトニアン（エネルギー関数：ミクロ世界のルール）
 - ▶ 例：調和振動子（バネの運動）

$$H(\mathbf{x}, \mathbf{p}) = \frac{1}{2}\mathbf{p}^2 + \frac{1}{2}\mathbf{x}^2$$

- ▶ 例：Ising模型（磁性体のモデル）

$$H(\mathbf{x}) = \frac{1}{2}\mathbf{x}^T J \mathbf{x}$$

- ▶ 確率分布による記述

- ▶ Gibbs-Boltzmann分布

$$P(\mathbf{x}) = \frac{1}{Z} \exp \left(-\frac{H(\mathbf{x})}{T} \right)$$

- ▶ 分配関数

$$Z = \int d\mathbf{x} \exp \left(-\frac{H(\mathbf{x})}{T} \right)$$

- ▶ 分配関数の計算命!
 - ▶ 等エネルギー面にある状態の数を先に勘定する.

$$Z = \int dE \int d\mathbf{x} \delta(H(\mathbf{x}) - E) \exp\left(-\frac{E}{T}\right)$$

▶ 分配関数の計算命!

- ▶ 等エネルギー面にある状態の数を先に勘定する.

$$Z = \int dE \int d\mathbf{x} \delta(H(\mathbf{x}) - E) \exp\left(-\frac{E}{T}\right)$$

- ▶ エントロピーの登場 $S(E) = \log \int d\mathbf{x} \delta(H(\mathbf{x}) - E)$

$$Z = \int dE \exp(S(E) - E/T)$$

▶ 分配関数の計算命!

- ▶ 等エネルギー一面にある状態の数を先に勘定する.

$$Z = \int dE \int d\mathbf{x} \delta(H(\mathbf{x}) - E) \exp\left(-\frac{E}{T}\right)$$

- ▶ エントロピーの登場 $S(E) = \log \int d\mathbf{x} \delta(H(\mathbf{x}) - E)$

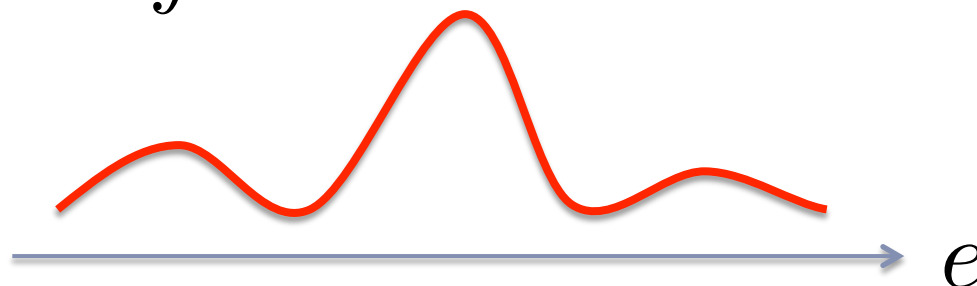
$$Z = \int dE \exp(S(E) - E/T)$$

- ▶ とてつもない自由度の数であるということを思い出す. (熱力学的極限)

$$\mathbf{x} \in \mathbb{R}^N \quad N \rightarrow \infty$$

- ▶ 一気にマクロの世界へ!
- ▶ 積分の意味：全領域の足し上げ！

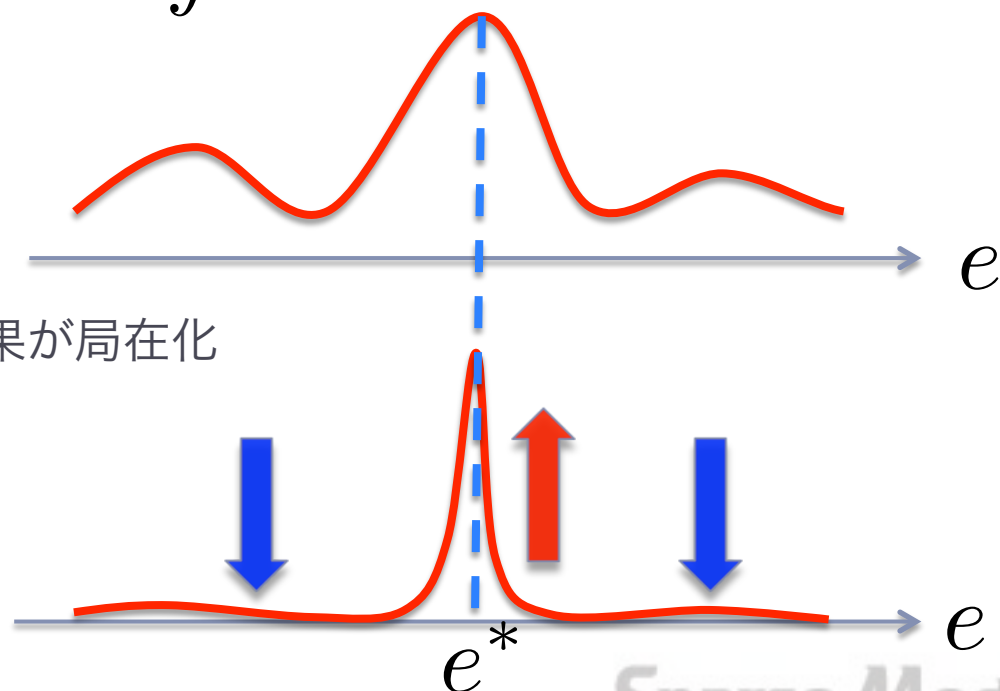
$$Z = N \int de \exp \{ N (s(e) - e/T) \}$$



- ▶ 一気にマクロの世界へ!
- ▶ 積分の意味：全領域の足し上げ!

$$Z = N \int de \exp \{ N (s(e) - e/T) \}$$

- ▶ Nを大きくすると…効果が局在化



- ▶ 一気にマクロの世界へ!

- ▶ 鞍点評価

$$Z \sim \exp \{ N (s(e^*) - e^*/T) \}$$

- ▶ 自由エネルギーの登場

$$f = -\frac{T}{N} \log Z$$

- ▶ 一気にマクロの世界へ!

- ▶ 鞍点評価

$$Z \sim \exp \{ N (s(e^*) - e^*/T) \}$$

- ▶ 自由エネルギーの登場 $f = -\frac{T}{N} \log Z$

- ▶ 分配関数の評価を通じて、全ての必要な熱力学量が計算できる!

$$f = e^* - Ts(e^*)$$

「物質は、エネルギーが低く、エントロピーの大きい、平衡状態を実現する！」

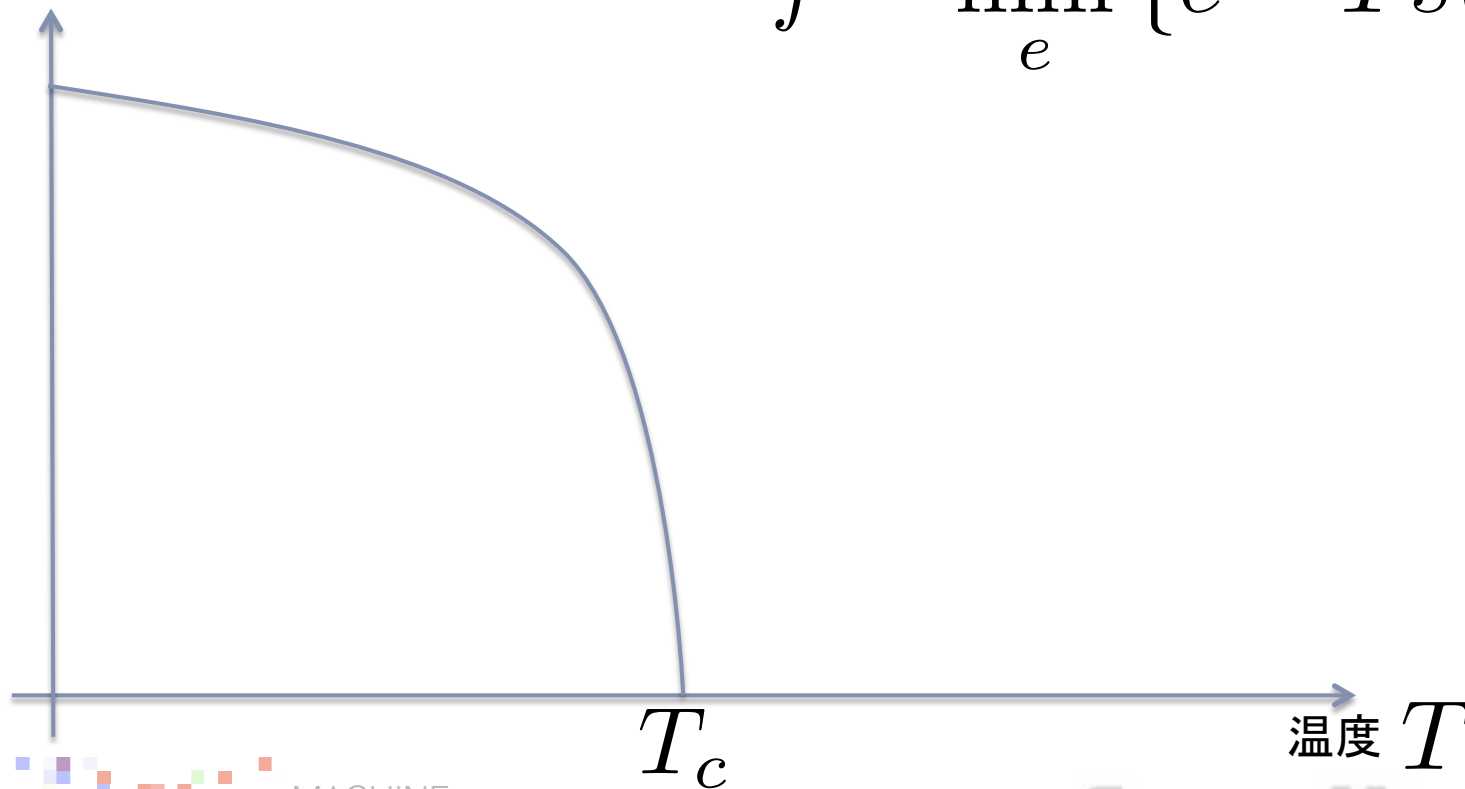
$$f = \min_e \{ e - Ts(e) \}$$

自由エネルギーの形状と相転移

- ▶ 磁性体の場合：Ising模型 $H(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T J \mathbf{x}$

$$f = \min_e \{e - Ts(e)\}$$

磁化 m



QUANTUM
ANNEALING



MACHINE
LEARNING

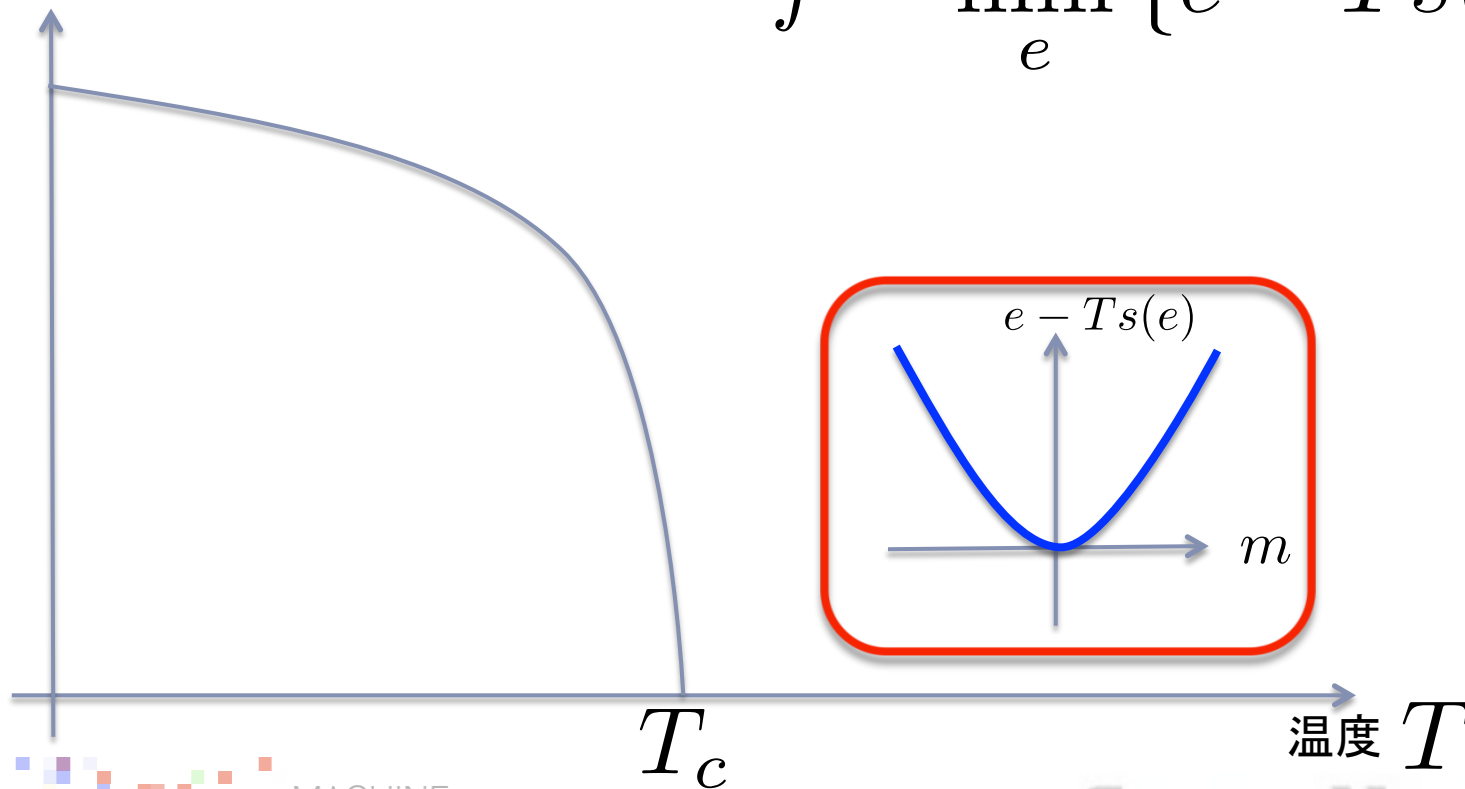
Sparse Modeling

自由エネルギーの形状と相転移

- ▶ 磁性体の場合：Ising模型 $H(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T J \mathbf{x}$

$$f = \min_e \{e - Ts(e)\}$$

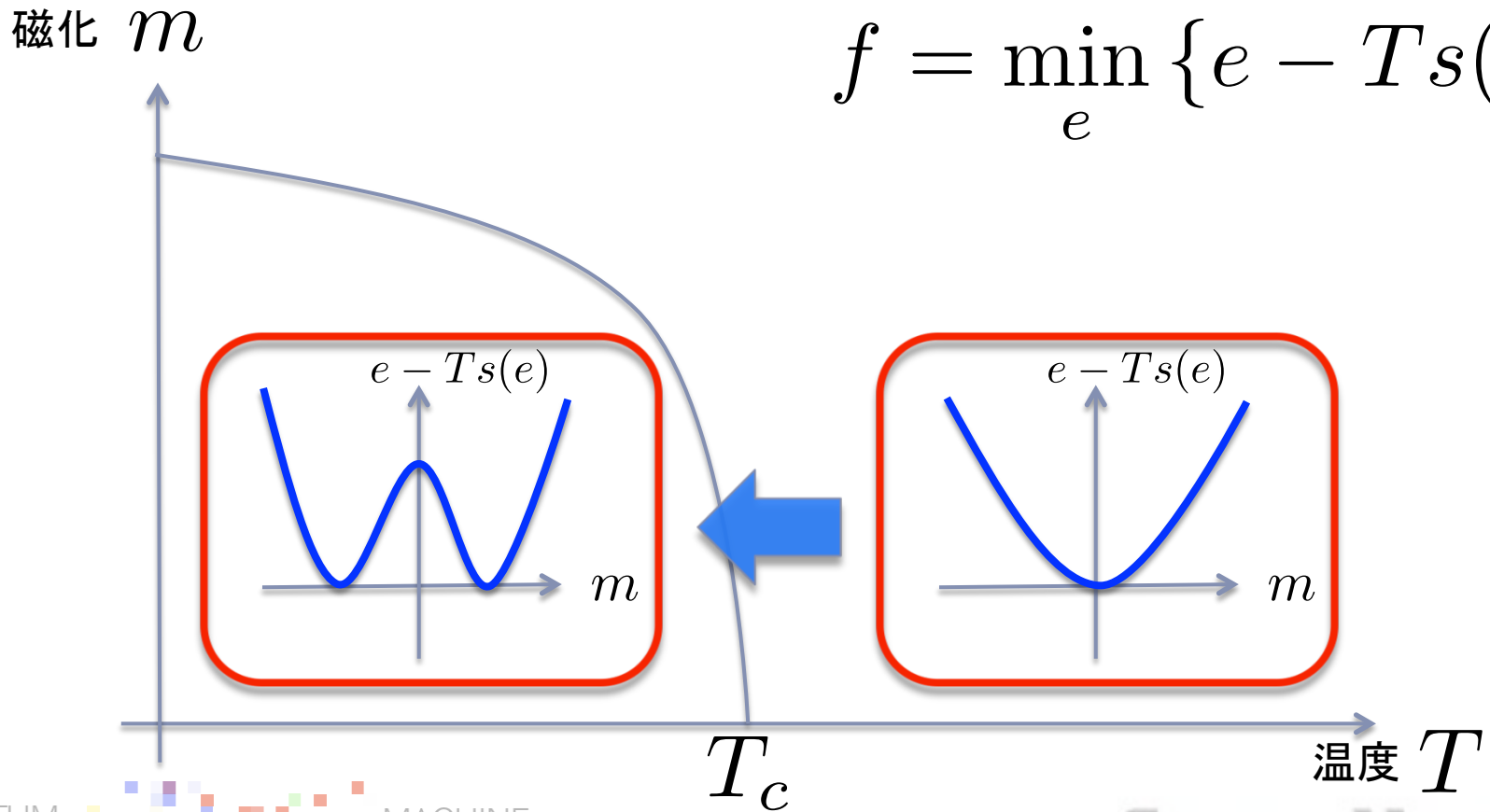
磁化 m



自由エネルギーの形状と相転移

- 磁性体の場合：Ising模型 $H(\mathbf{x}) = \frac{1}{2} \mathbf{x}^T J \mathbf{x}$

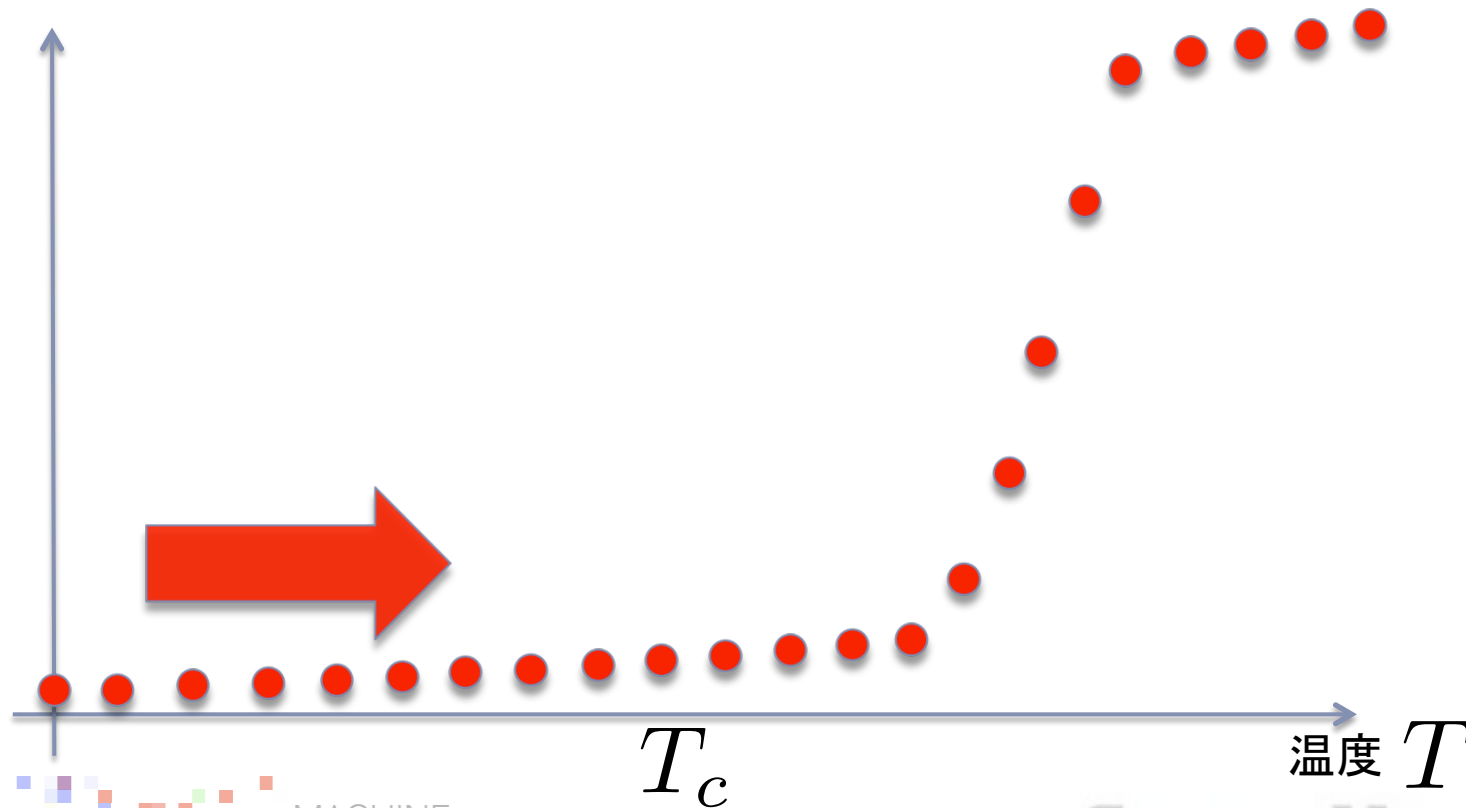
$$f = \min_e \{e - Ts(e)\}$$



自由エネルギーの形状と相転移

- ▶ 変わった磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

内部エネルギー e



QUANTUM
ANNEALING



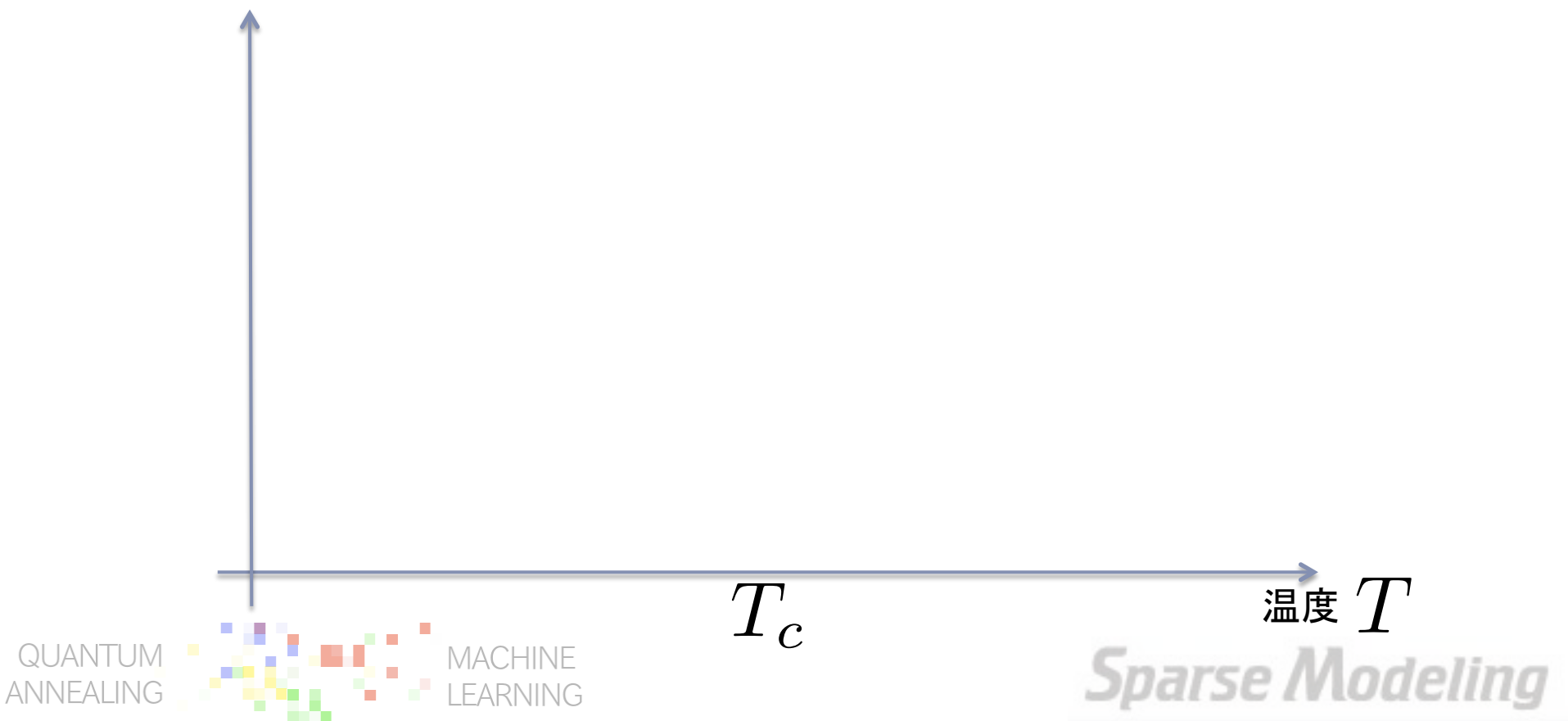
MACHINE
LEARNING

Sparse Modeling

自由エネルギーの形状と相転移

- ▶ 変わった磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

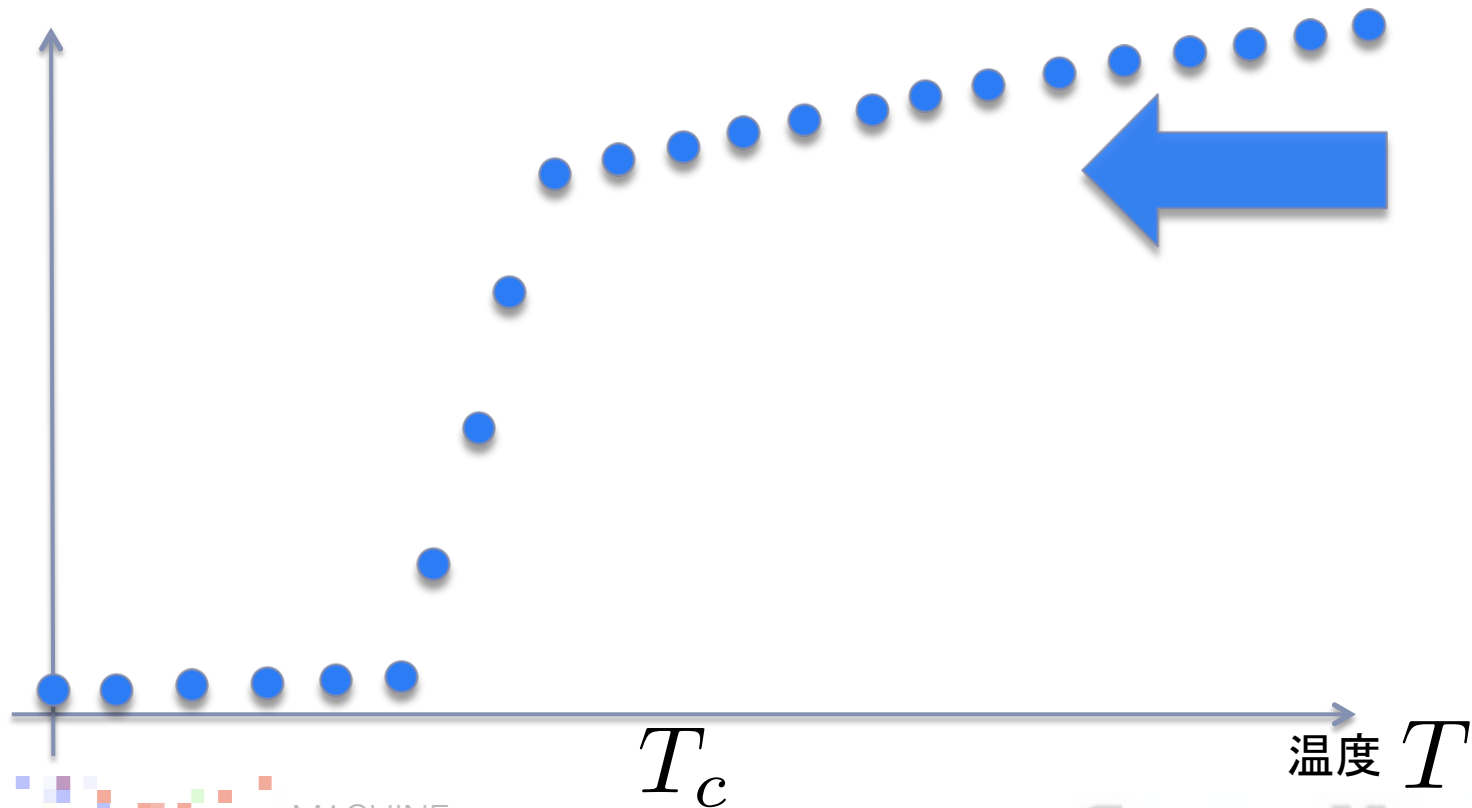
内部エネルギー e



自由エネルギーの形状と相転移

- ▶ 変わった磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

内部エネルギー e



QUANTUM
ANNEALING



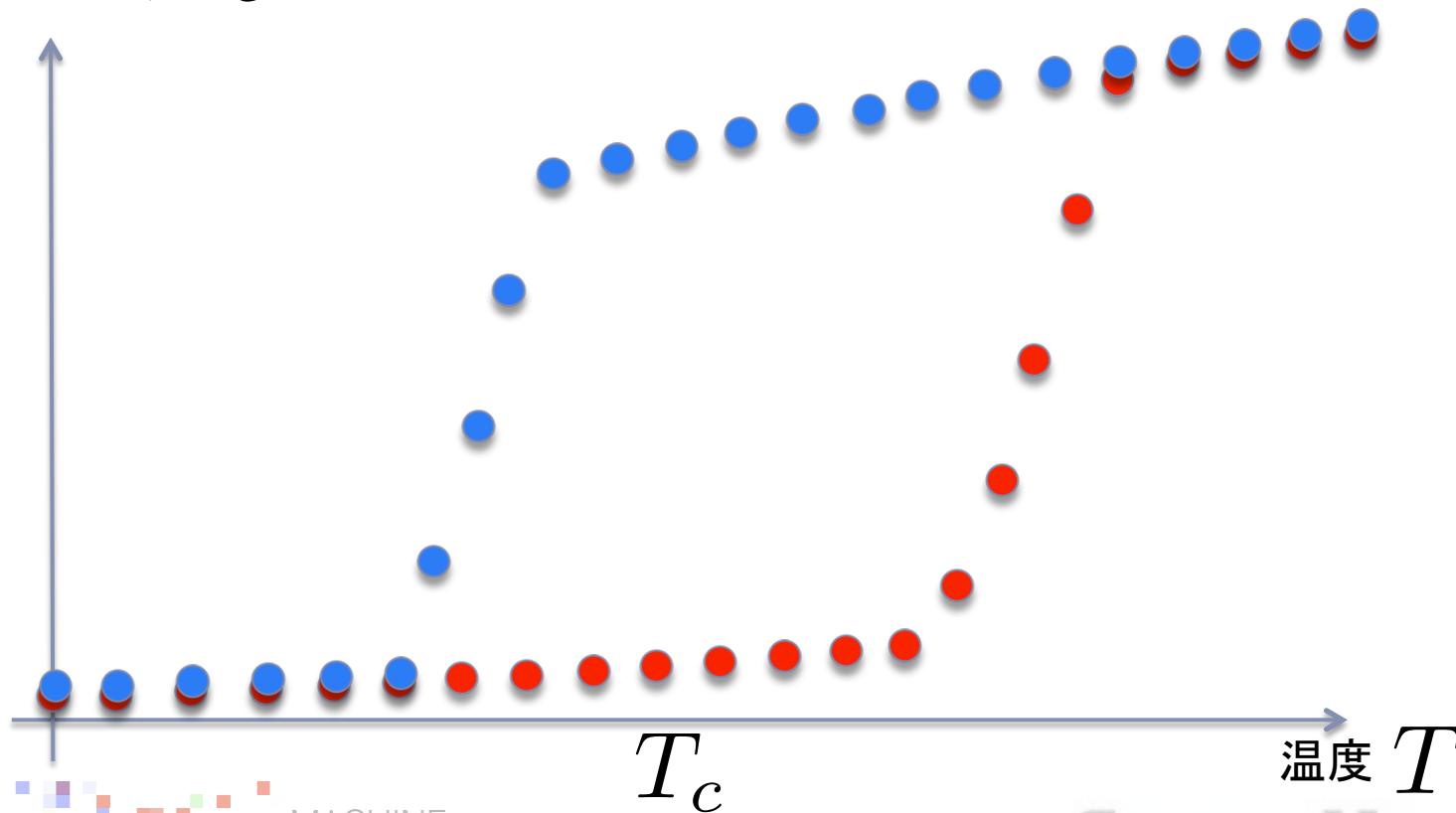
MACHINE
LEARNING

Sparse Modeling

自由エネルギーの形状と相転移

- ▶ 変わった磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

内部エネルギー e



QUANTUM
ANNEALING

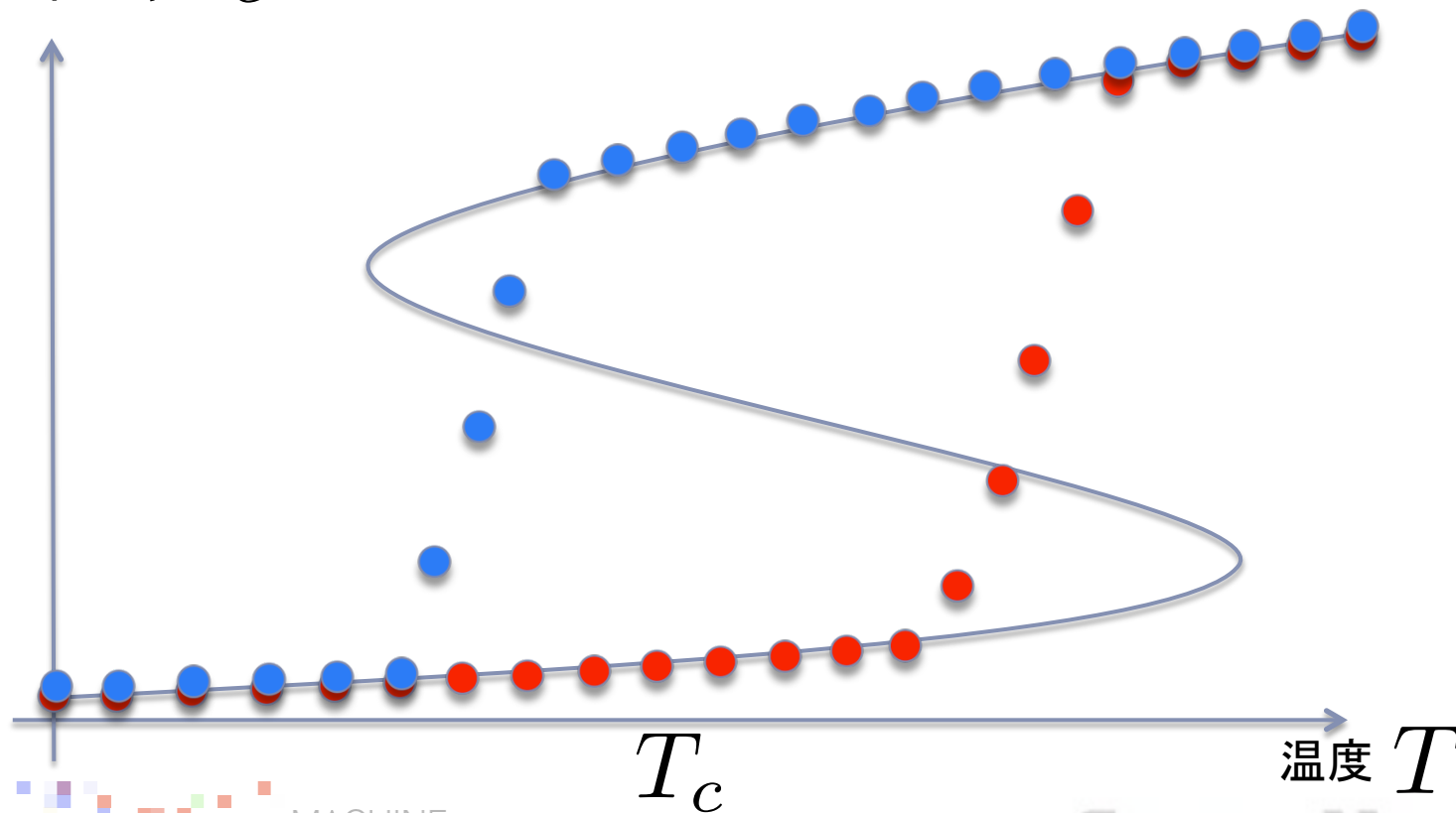
MACHINE
LEARNING

Sparse Modeling

自由エネルギーの形状と相転移

- ▶ 変わった磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

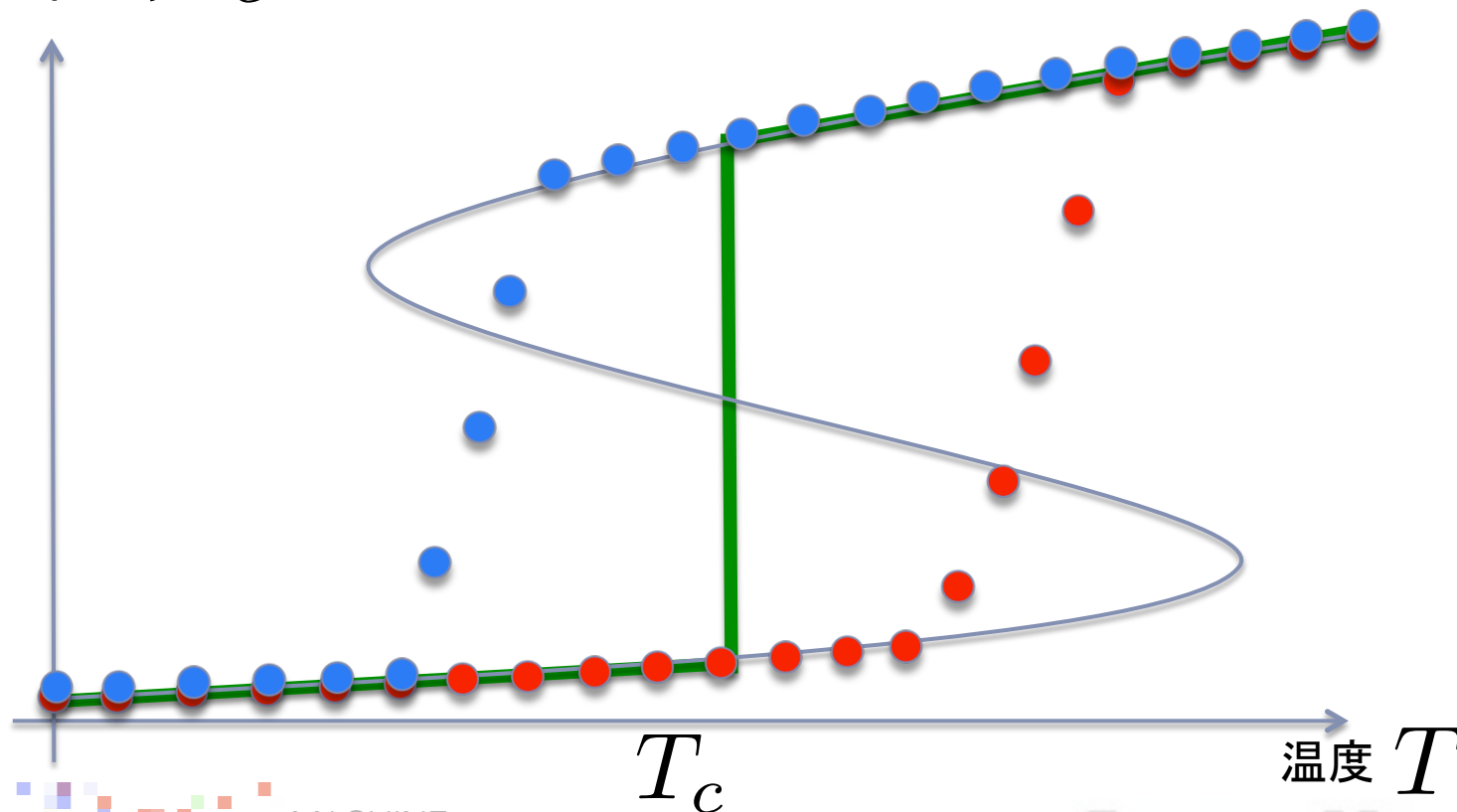
内部エネルギー e



自由エネルギーの形状と相転移

- ▶ 変わった磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

内部エネルギー e



QUANTUM
ANNEALING



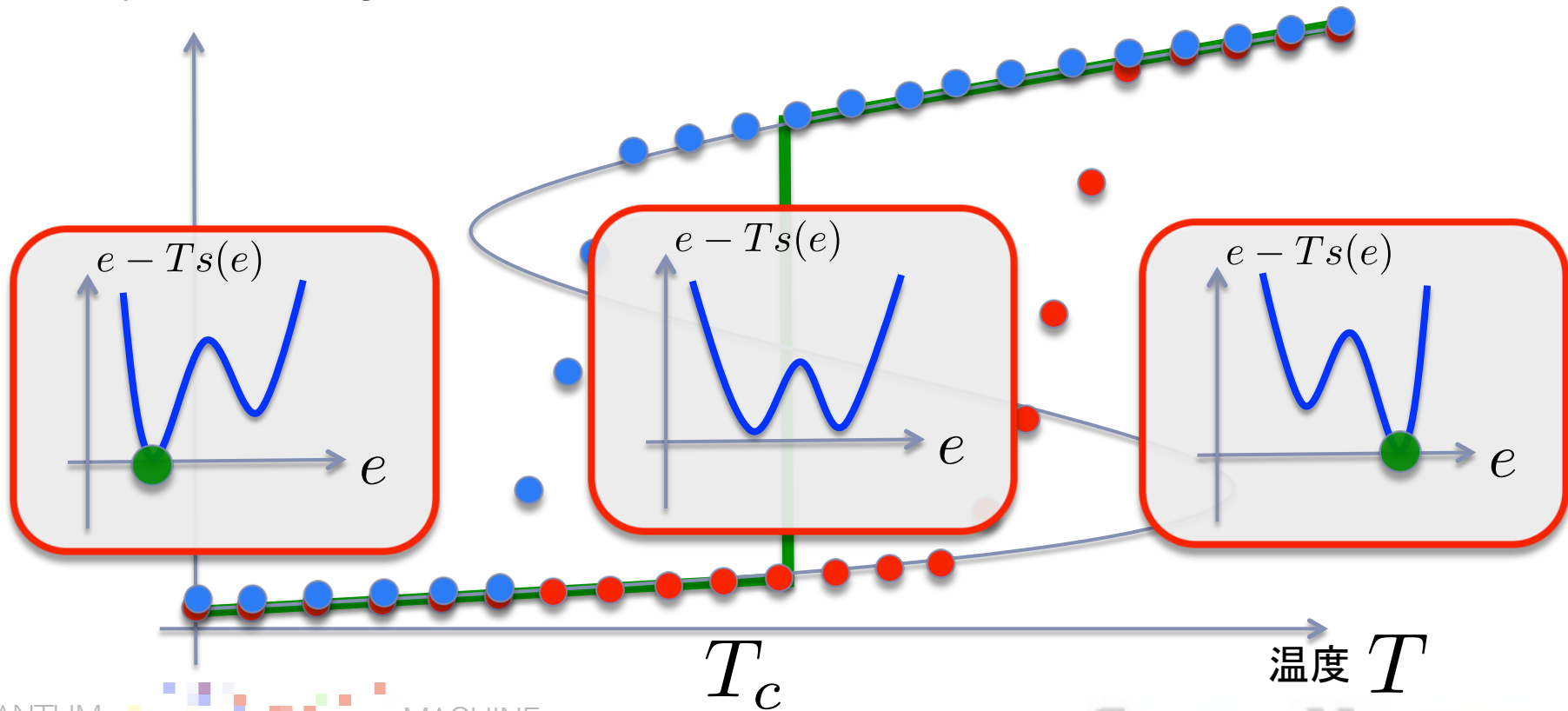
MACHINE
LEARNING

Sparse Modeling

自由エネルギーの形状と相転移

- 変った磁性体の場合：Potts模型 $f = \min_e \{e - Ts(e)\}$

内部エネルギー e



QUANTUM
ANNEALING



MACHINE
LEARNING

Sparse Modeling

- ▶ 基底追跡型の最適化問題の解析
 - ▶ どんな $\mathbf{x}$ が生成されるか？

$$P(\mathbf{x}|A, \mathbf{y}) \propto \exp(-\beta \|\mathbf{x}\|_1) \delta(\mathbf{y} - A\mathbf{x})$$

- ▶ 逆温度 β 無限大で、

$$\min_{\mathbf{x}} \|\mathbf{x}\|_1 \quad \text{s.t.} \quad \mathbf{y} = A\mathbf{x}$$

- ▶ 様々な A と $\mathbf{y}$ での評価を必要とするので、レプリカ法を利用する.

$$[\log Z(A, \mathbf{y})]_{A, \mathbf{y}} = \lim_{n \rightarrow 0} \frac{[Z^n(A, \mathbf{y})]_{A, \mathbf{y}} - 1}{n}$$

▶ 分配関数の評価

▶ nレプリカの系における多重積分の評価

$$[Z^n(A, \mathbf{y})]_{A, \mathbf{y}} = \left[\int \prod_{a=1}^n d\mathbf{x}_a \exp(-\beta \|\mathbf{x}_a\|_1) \delta(\mathbf{y} - A\mathbf{x}^a) \right]_{A, \mathbf{y}}$$

▶ 罰金法

$$\delta(\mathbf{y} - A\mathbf{x}^a) = \exp\left(-\frac{1}{2\lambda} \|\mathbf{y} - A\mathbf{x}\|_2^2\right) \quad (\lambda \rightarrow +0)$$

▶ 仮定

- ノイズなし観測過程 $\mathbf{y} = A\mathbf{x}^0$
- Aは各成分ガウスランダム行列 (平均0、分散1/N)
- $\mathbf{x}^0$ は各成分

$$P_0(x) = (1 - \rho)\delta(x) + \rho \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2}x^2\right)$$

▶ 自由エネルギー (nの1次部分)

▶ レプリカ対称性の下、秩序パラメータ Q, χ, m と共役パラメータの鞍点評価

$$-f = \max_{Q, \tilde{Q}} \left\{ -\frac{\alpha}{2\chi} (\rho - 2m + Q) + \frac{1}{2} (Q\tilde{Q} - \chi\tilde{\chi}) - m\tilde{m} \right. \\ \left. + \frac{1-\rho}{\tilde{Q}} G(\tilde{\chi}) + \frac{\rho}{\tilde{Q}} G(\tilde{\chi} + \tilde{m}^2) \right\}$$

▶ 関数の定義

$$G(a) = (a+1)H\left(\frac{1}{\sqrt{a}}\right) - \sqrt{\frac{a}{2\pi}} \exp\left(-\frac{1}{2a}\right)$$

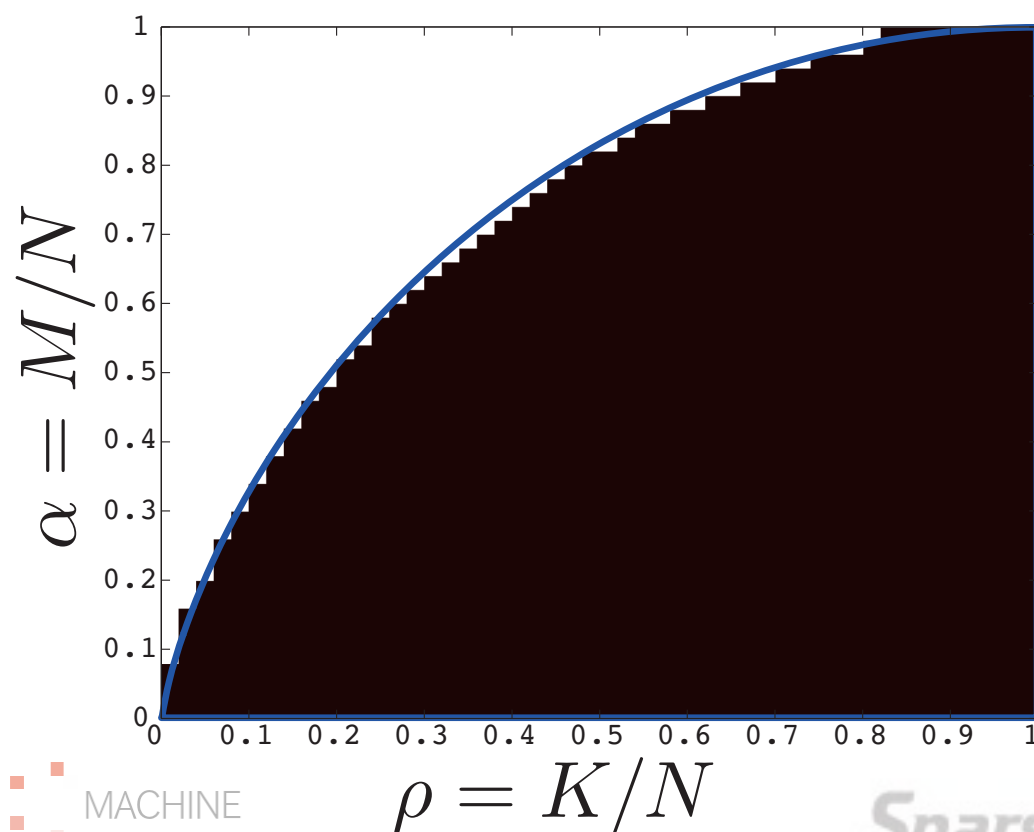
$$H(x) = \int_x^\infty Dz$$

鞍点方程式の解

相転移

性能指標

$$\text{MSE} = \frac{1}{N} \|\mathbf{x} - \mathbf{x}_0\|_2^2$$



▶ 基底追跡型の最適化問題の解析

▶ どんな $\mathbf{x}$ が生成されるか？

$$P(\mathbf{x}|A, \mathbf{y}) \propto \exp(-\beta \|\mathbf{x}\|_1) \delta(\mathbf{y} - A\mathbf{x})$$

▶ 罰金法

$$\delta(\mathbf{y} - A\mathbf{x}^a) = \exp\left(-\frac{1}{2\lambda} \|\mathbf{y} - A\mathbf{x}\|_2^2\right) (\lambda \rightarrow +0)$$

▶ ラグランジュ未定乗数法

$$\delta(\mathbf{y} - A\mathbf{x}^a) = \exp(i\mathbf{h}^T (\mathbf{y} - A\mathbf{x}))$$

▶ 拡張ラグランジュ法

$$\delta(\mathbf{y} - A\mathbf{x}^a) = \exp\left(i\mathbf{h}^T (\mathbf{y} - A\mathbf{x}) - \frac{1}{2\lambda} \|\mathbf{y} - A\mathbf{x}\|_2^2\right)$$

▶ どれも等価な結果を導く

▶ LASSO型最適化問題

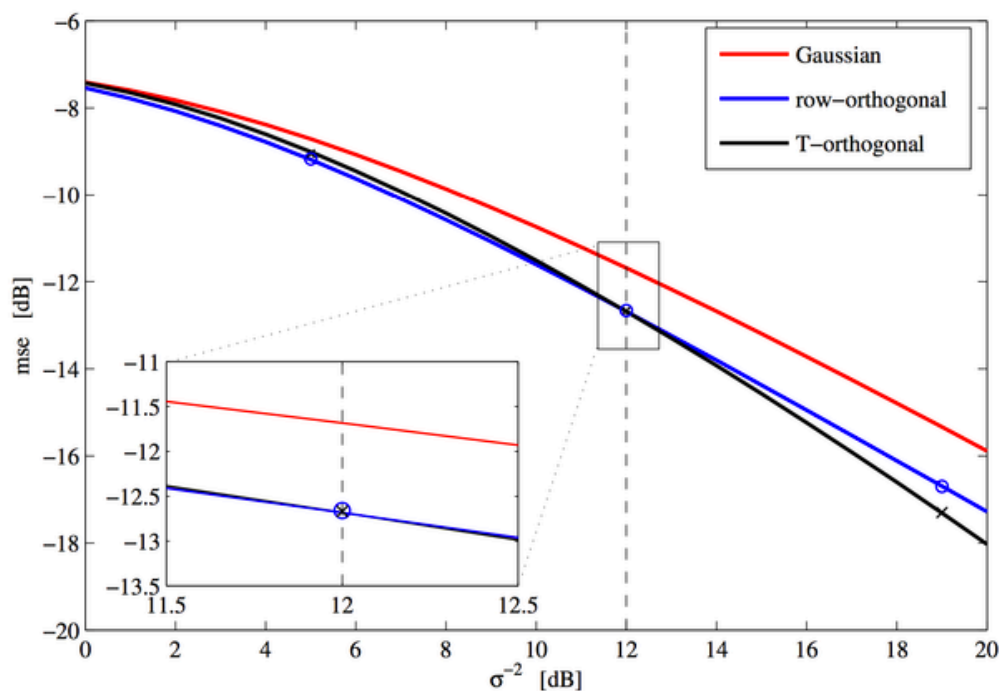
$$P(\mathbf{x}|A, \mathbf{y}) \propto \exp \left(-\beta \|\mathbf{x}\|_1 - \frac{\beta}{2\lambda} \|\mathbf{y} - A\mathbf{x}\|_2^2 \right)$$

▶ ノイズ有り圧縮センシング

$$\mathbf{y} = A\mathbf{x}_0 + \sigma_0 \mathbf{w}$$

▶ 観測行列の違いによる性能比較

- ▶ ガウスランダム
- ▶ 直交変換
- ▶ T-直交

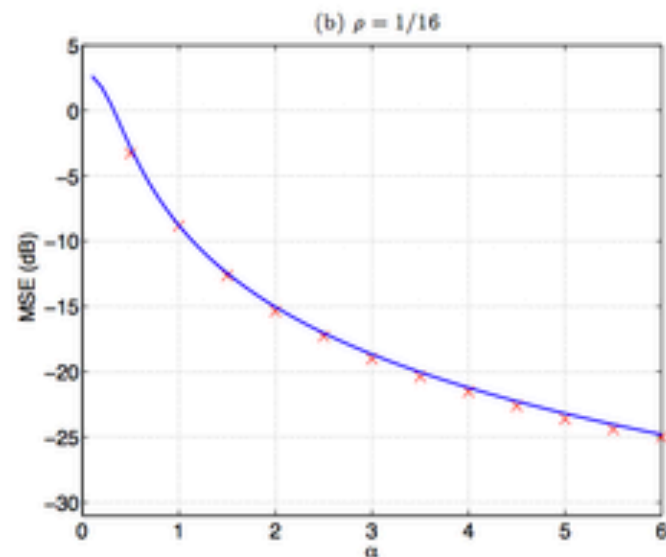
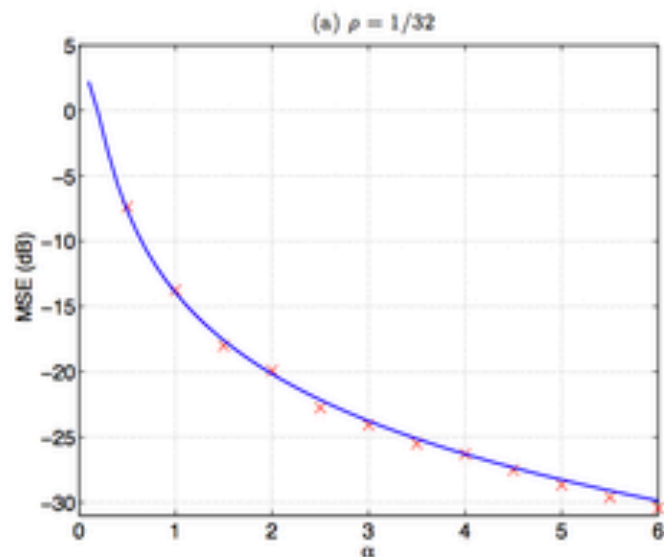


- ▶ 1-bit 圧縮センシング

$$\mathbf{y} = \text{sign} (A\mathbf{x})$$

- ▶ 符号情報だけ送られる (情報通信に利用)

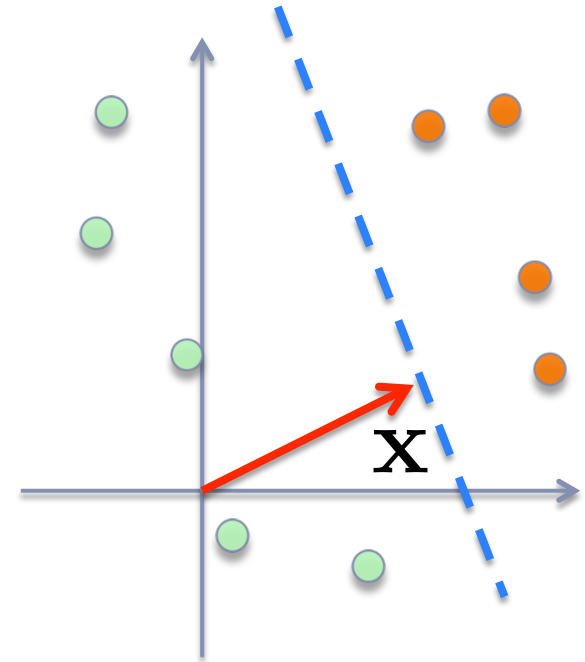
$$P(\mathbf{x}|A, \mathbf{y}) \propto \exp (-\beta \|\mathbf{x}\|_1) \delta \left(\|\mathbf{x}\|_2^2 - 1 \right) \Theta (\mathbf{y} A \mathbf{x})$$



- ▶ 各種ニューラルネットワーク問題
 - ▶ ガードナー容量問題

$$\mathbf{y} = \text{sign} (A\mathbf{x})$$

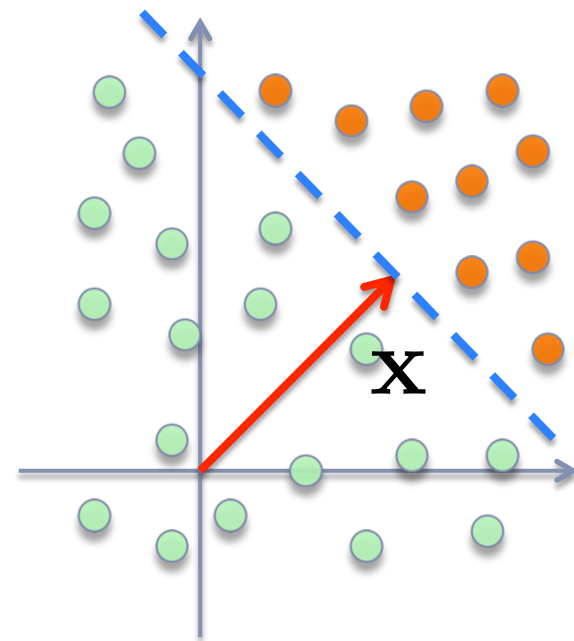
$$P(\mathbf{x}|A, \mathbf{y}) \propto \delta \left(\|\mathbf{x}\|_2^2 - 1 \right) \Theta (\mathbf{y} A\mathbf{x})$$



- ▶ 各種ニューラルネットワーク問題
 - ▶ ガードナー容量問題

$$\mathbf{y} = \text{sign} (A\mathbf{x})$$

$$P(\mathbf{x}|A, \mathbf{y}) \propto \delta \left(\|\mathbf{x}\|_2^2 - 1 \right) \Theta (\mathbf{y} A\mathbf{x})$$

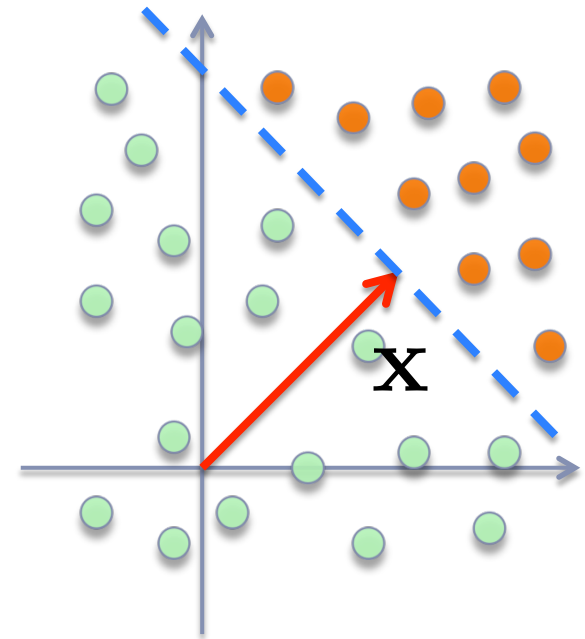


- ▶ 各種ニューラルネットワーク問題
 - ▶ ガードナー容量問題

$$\mathbf{y} = \text{sign} (A\mathbf{x})$$

$$P(\mathbf{x}|A, \mathbf{y}) \propto \delta \left(\|\mathbf{x}\|_2^2 - 1 \right) \Theta (\mathbf{y} A\mathbf{x})$$

- ▶ サポートベクトルマシン
- ▶ 多層ニューラルネットワークも可能

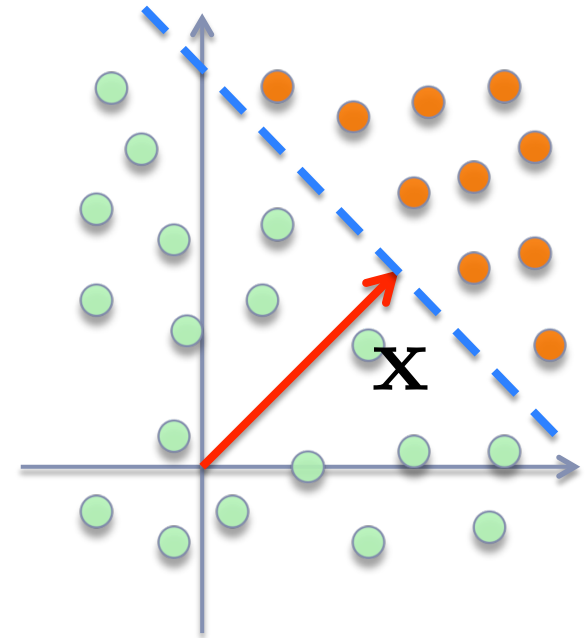


- ▶ 各種ニューラルネットワーク問題
 - ▶ ガードナー容量問題

$$\mathbf{y} = \text{sign} (A\mathbf{x})$$

$$P(\mathbf{x}|A, \mathbf{y}) \propto \delta \left(\|\mathbf{x}\|_2^2 - 1 \right) \Theta (\mathbf{y} A\mathbf{x})$$

- ▶ サポートベクトルマシン
- ▶ 多層ニューラルネットワークも可能
- ▶ 深層学習は??



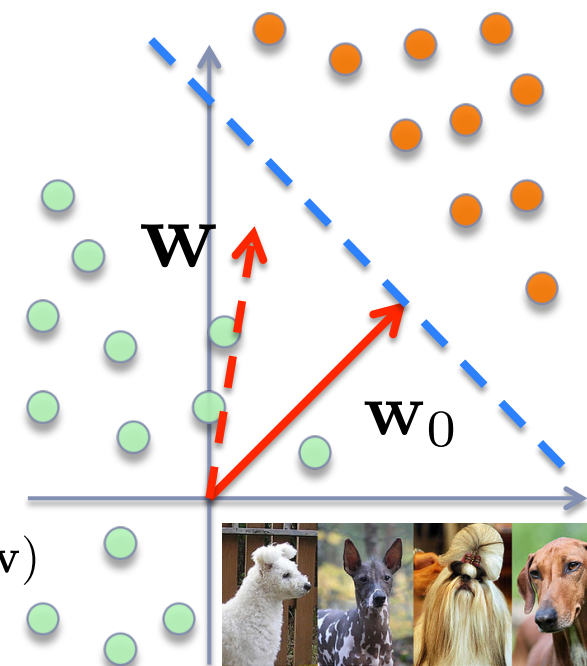
- ▶ 教師無し学習と教師有り学習の効果を見る
 - ▶ 手で解析できるように極端に単純化
 - ▶ 単純パーセプトロン+大量の高次元データ
 - ▶ 入力データに構造があるところがポイント
 - ▶ 教師無し学習でも意味がある情報利得を仮定

$$P(\mathbf{x}_\mu, y_\mu | \mathbf{w}) \propto \Theta \left(\frac{y_\mu}{\sqrt{N}} \mathbf{x}^T \mathbf{w} - h \right)$$

$$P(\mathbf{x}_\mu | \mathbf{w}) \propto \sum_{y_\mu} \Theta \left(\frac{y_\mu}{\sqrt{N}} \mathbf{x}^T \mathbf{w} - h \right)$$

- ▶ ベイズ推定

$$P(\mathbf{w} | X, \mathbf{y}) = \frac{1}{Z} \prod_{\mu=1}^L P(\mathbf{x}_\mu, y_\mu | \mathbf{w}) \prod_{\mu=L+1}^{L+U} P(\mathbf{x}_\mu | \mathbf{w}) P(\mathbf{w})$$



引用: wikipedia

▶ 教師無し学習と教師有り学習の効果を見る

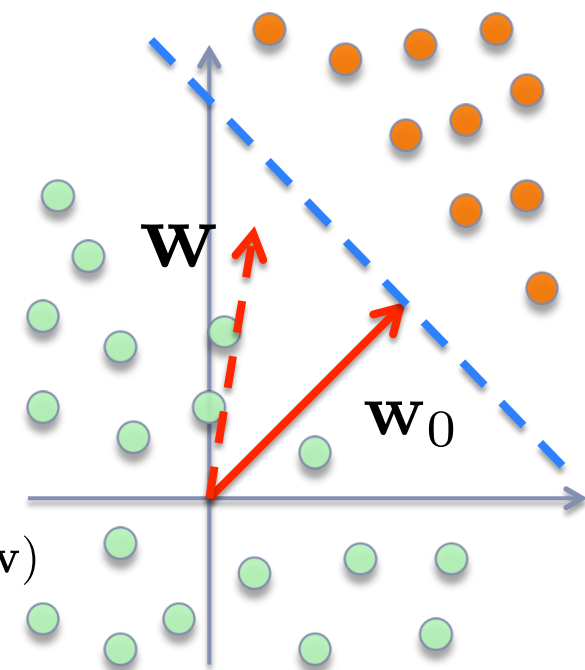
- ▶ 手で解析できるように極端に単純化
 - ▶ 単純パーセプトロン+大量の高次元データ
- ▶ 入力データに構造があるところがポイント
 - ▶ 教師無し学習でも意味がある情報利得を仮定

$$P(\mathbf{x}_\mu, y_\mu | \mathbf{w}) \propto \Theta \left(\frac{y_\mu}{\sqrt{N}} \mathbf{x}^T \mathbf{w} - h \right)$$

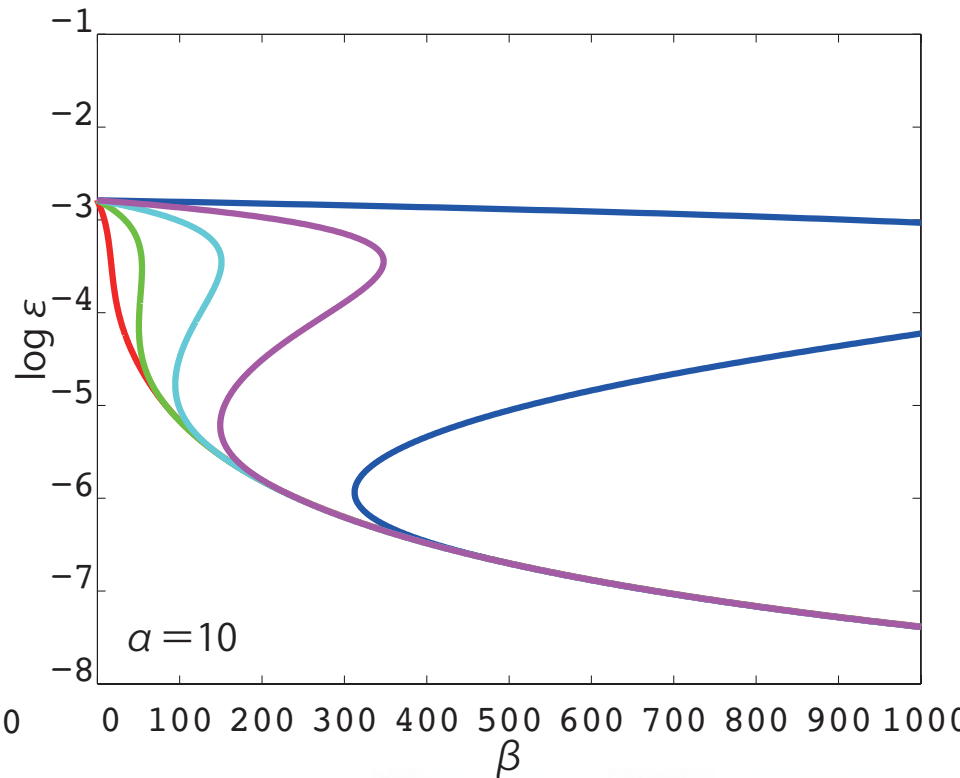
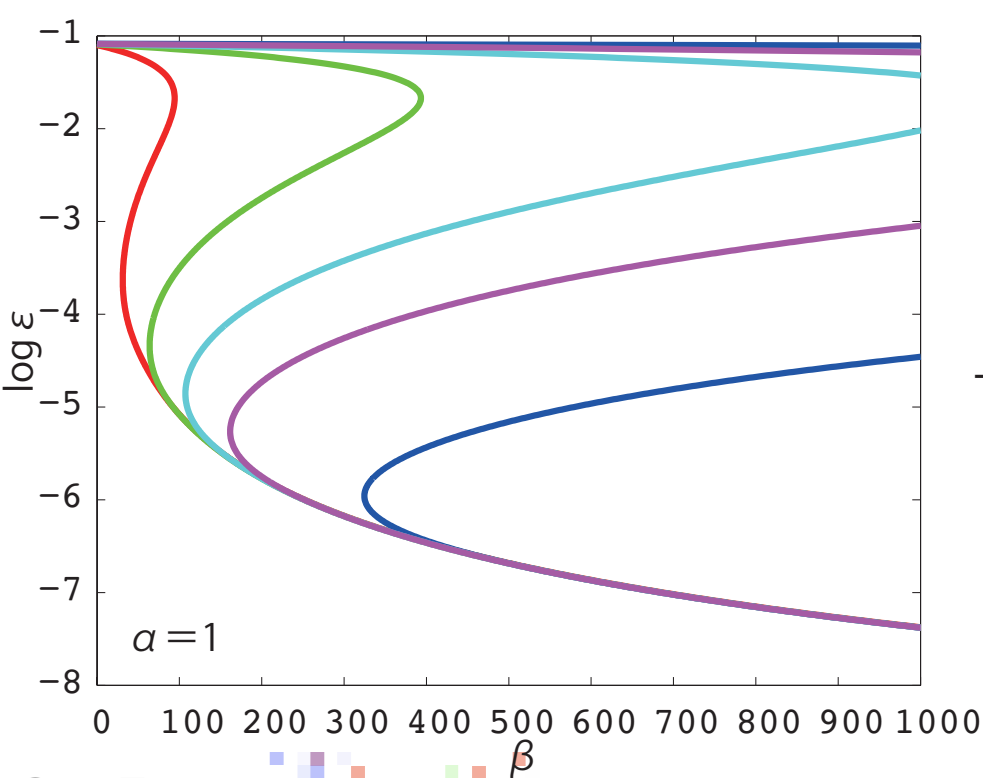
$$P(\mathbf{x}_\mu | \mathbf{w}) \propto \sum_{y_\mu} \Theta \left(\frac{y_\mu}{\sqrt{N}} \mathbf{x}^T \mathbf{w} - h \right)$$

▶ ベイズ推定

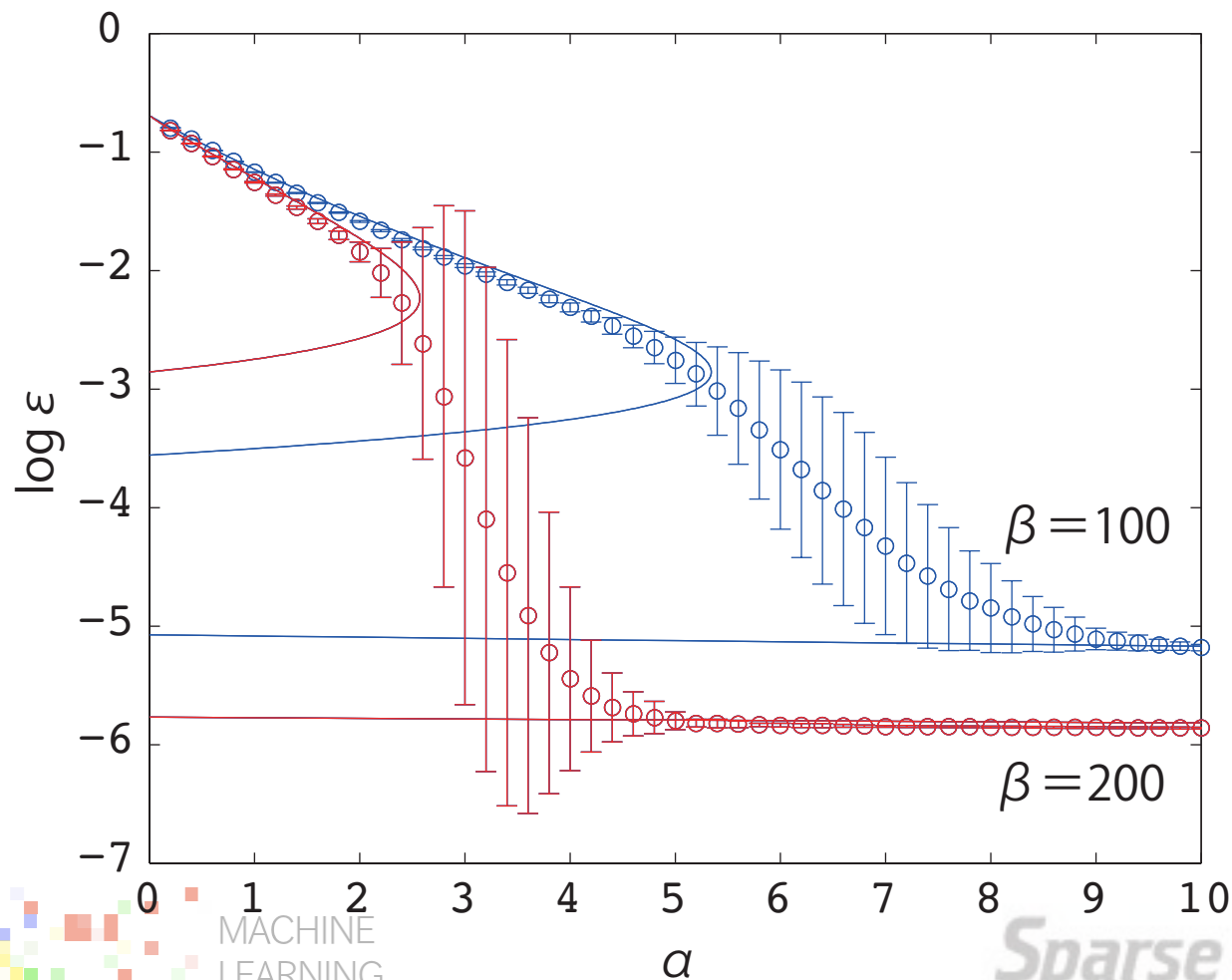
$$P(\mathbf{w} | X, \mathbf{y}) = \frac{1}{Z} \prod_{\mu=1}^L P(\mathbf{x}_\mu, y_\mu | \mathbf{w}) \prod_{\mu=L+1}^{L+U} P(\mathbf{x}_\mu | \mathbf{w}) P(\mathbf{w})$$



- ▶ 最尤法で学習できた場合の性能評価
 - ▶ 汎化誤差の解析結果
 - ▶ 赤、緑、水、紫、青とマージンが小さくなり識別が難しい場合を検討

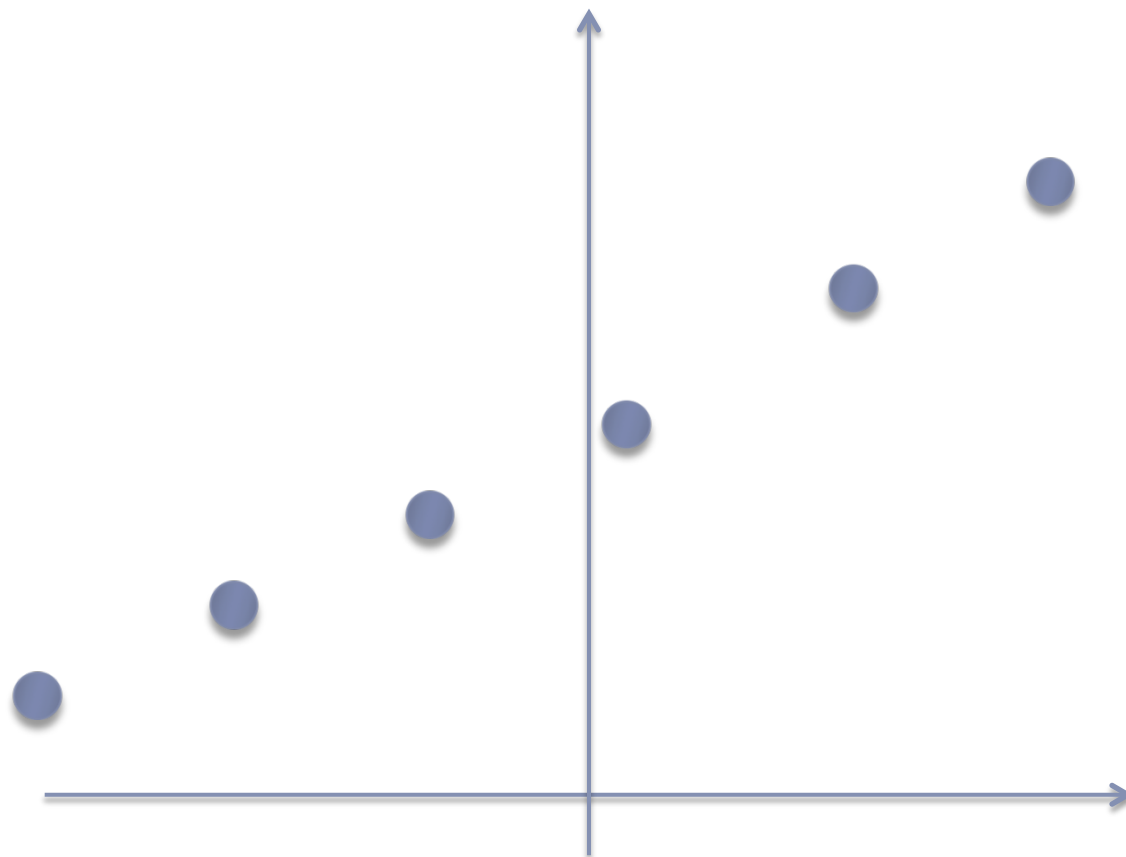


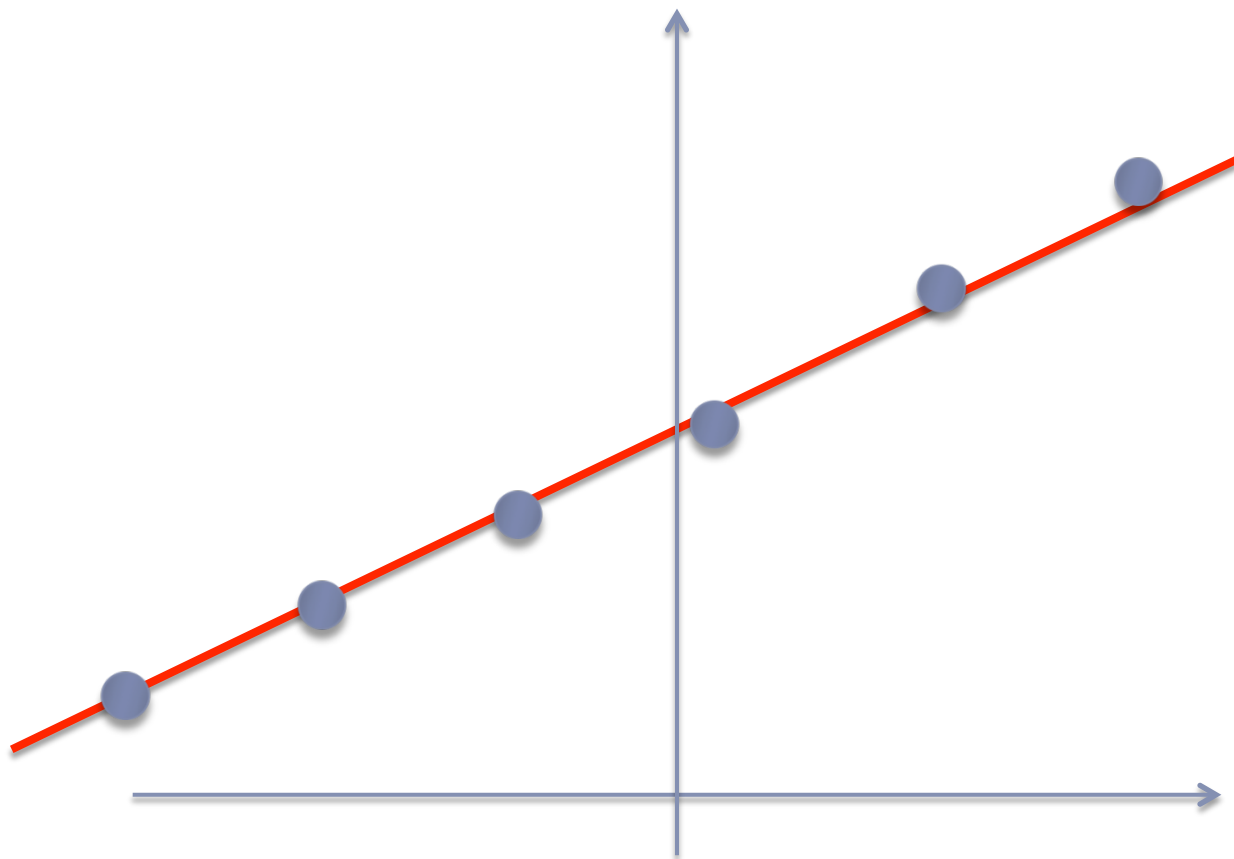
- 教師無し学習(β)と教師あり学習(α)の問題数における相転移

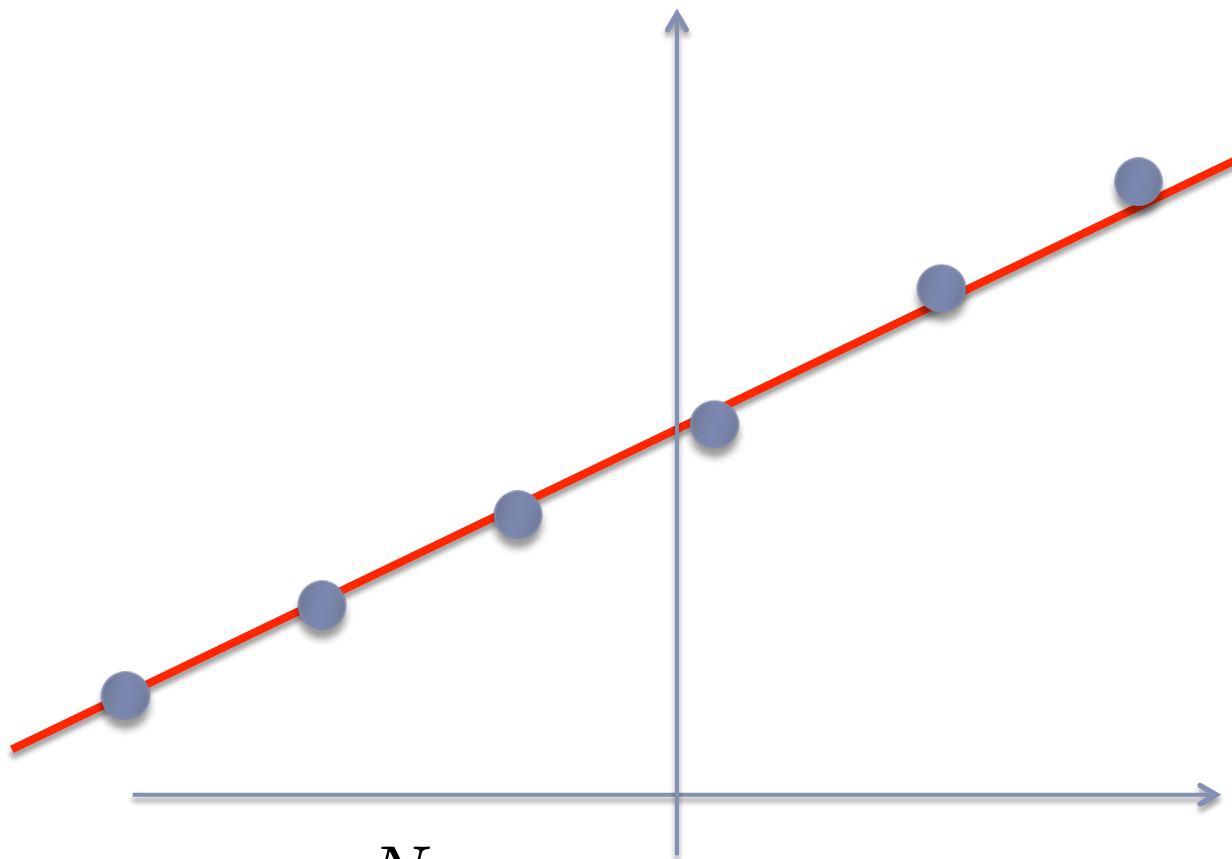


ボルツマン機械学習編

基本は関数の自律的フィッティングです



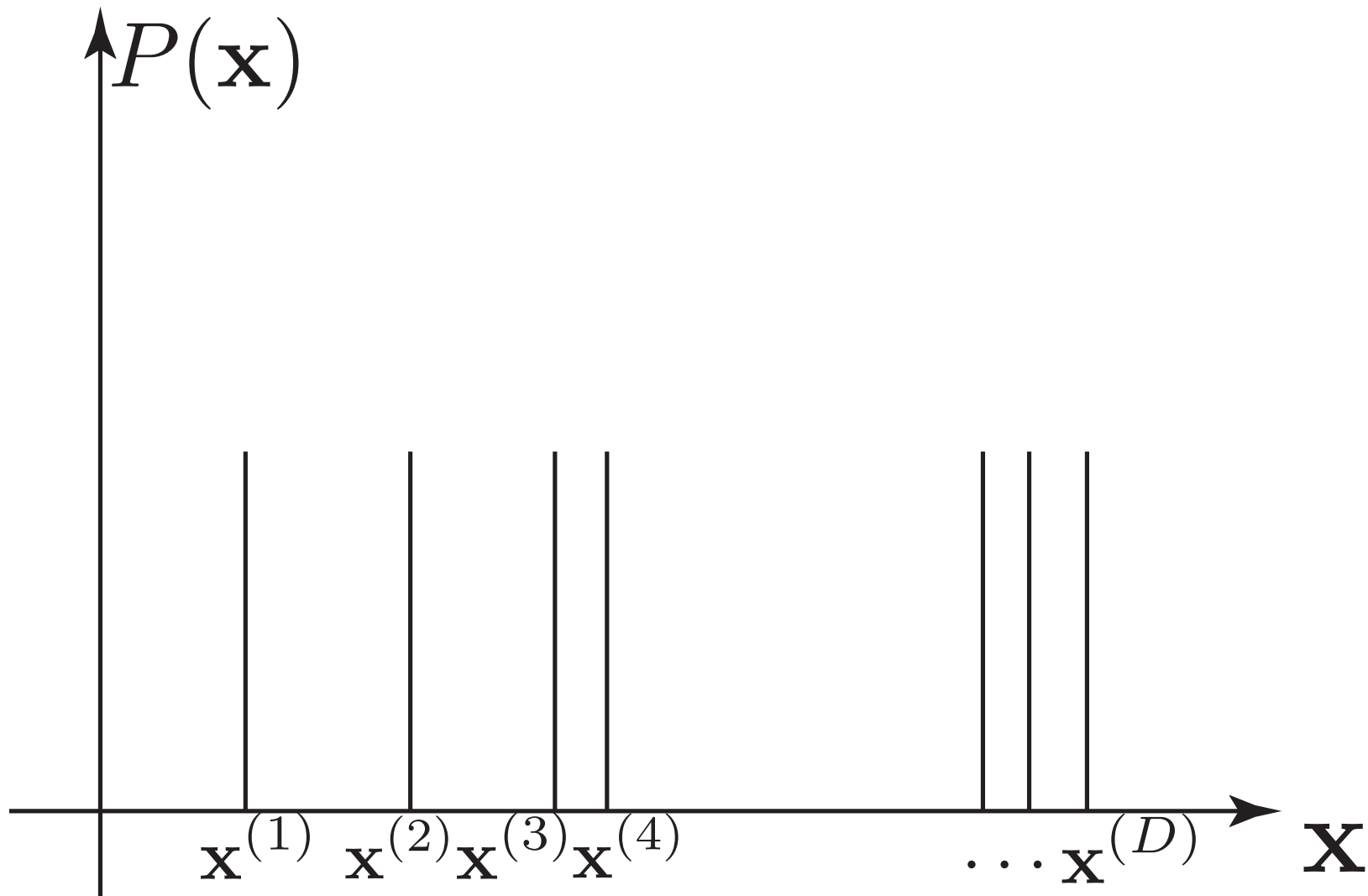


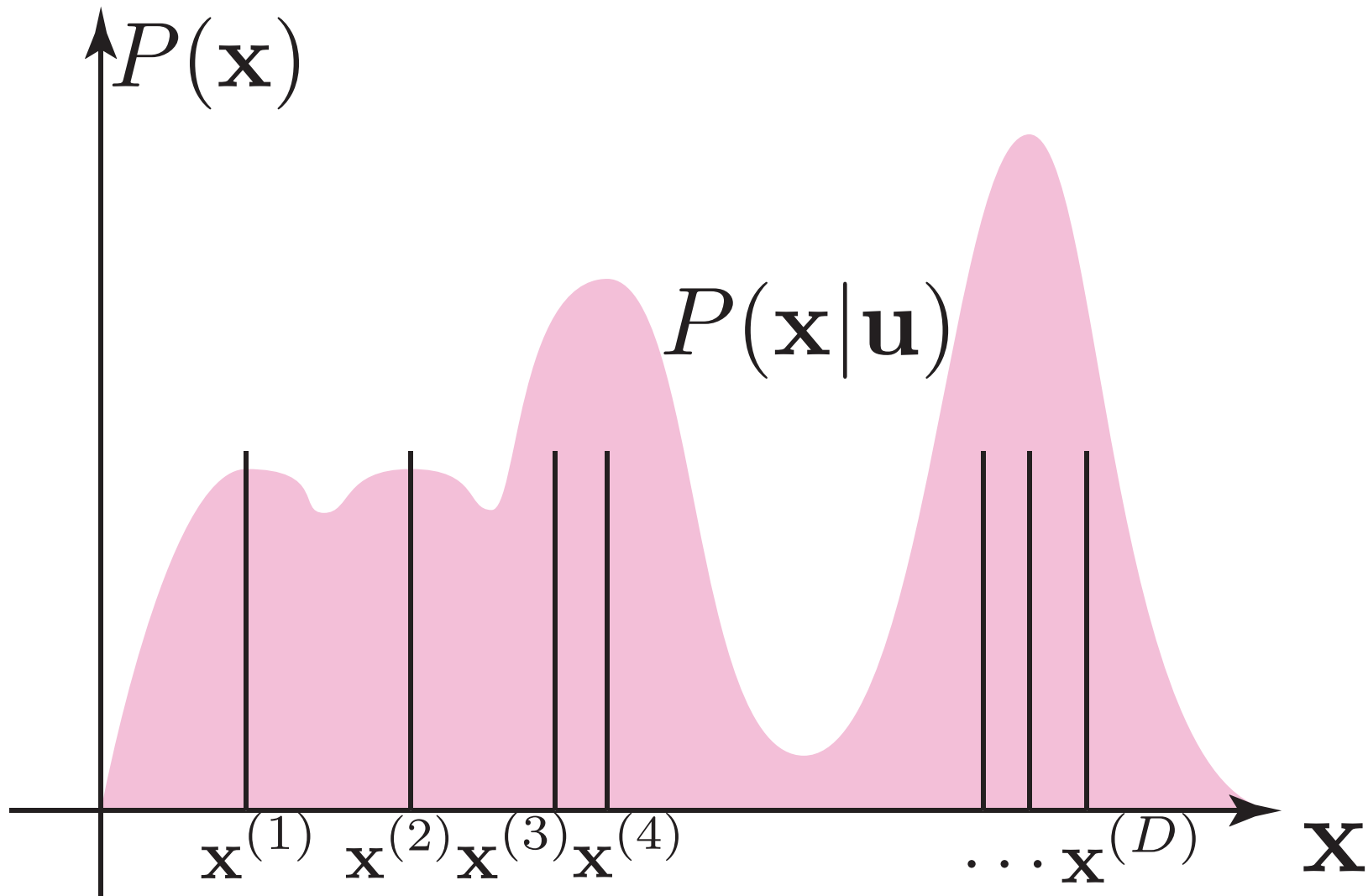


$$\min_{a,b} \sum_{i=1}^N \{y_i - (ax_i + b)\}^2$$

多くのデータからその素性を明らかにする営み
機械学習

多くのデータから確率分布を明らかにする
ボルツマン機械学習





統計的性質
平均？分散？

統計的性質
平均？分散？
分布を知りたい

- ▶ 多数の数値列、データからの分布の推定
 - ▶ 生成モデルの存在の仮定

$$P(\mathbf{x}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \{ -E(\mathbf{x}|\mathbf{u}) \}$$

- ▶ 例：Ising模型

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N \sum_{j \in \partial i} J_{ij} x_i x_j + \sum_{i=1}^N h_i x_i$$

- ▶ 多数の数値列、データからの分布の推定
 - ▶ 生成モデルの存在の仮定

$$P(\mathbf{x}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \{ -E(\mathbf{x}|\mathbf{u}) \}$$

- ▶ 例：Ising模型

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N \sum_{j \in \partial i} J_{ij} x_i x_j + \sum_{i=1}^N h_i x_i$$

- ▶ 利用例

- ▶ 相関検出

- カンニング検出、ニューロンシグナル、因子推定

- ▶ 多数の数値列、データからの分布の推定
 - ▶ 生成モデルの存在の仮定

$$P(\mathbf{x}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \{ -E(\mathbf{x}|\mathbf{u}) \}$$

- ▶ 例：Ising模型

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N \sum_{j \in \partial i} J_{ij} x_i x_j + \sum_{i=1}^N h_i x_i$$

- ▶ 利用例

- ▶ 相関検出
 - カンニング検出、ニューロンシグナル、因子推定
- ▶ ハミルトニアン推定
 - 物理モデルの正当化、パラメータ推定

- ▶ 多数の数値列、データからの分布の推定
 - ▶ 生成モデルの存在の仮定

$$P(\mathbf{x}|\mathbf{u}) = \frac{1}{Z(\mathbf{u})} \exp \{ -E(\mathbf{x}|\mathbf{u}) \}$$

- ▶ 例：Ising模型

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N \sum_{j \in \partial i} J_{ij} x_i x_j + \sum_{i=1}^N h_i x_i$$

- ▶ 利用例

- ▶ 相関検出
 - カンニング検出、ニューロンシグナル、因子推定
- ▶ ハミルトニアン推定
 - 物理モデルの正当化、パラメータ推定
- ▶ 特徴抽出

- ▶ 職人技の学習、自然の学習





ボルツマン機械学習の基礎

- ▶ データとモデルがどれだけ一致しているか？

- ▶ Kullback-Leibler divergence

$$D_{\text{KL}}(P|Q) = \int d\mathbf{x} P(\mathbf{x}) \log \left(\frac{P(\mathbf{x})}{Q(\mathbf{x})} \right)$$

- ▶ 非負
- ▶ 三角不等式を満たさない、非対称

- ▶ データとモデルがどれだけ一致しているか？

- ▶ Kullback-Leibler divergence

$$D_{\text{KL}}(P|Q) = \int d\mathbf{x} P(\mathbf{x}) \log \left(\frac{P(\mathbf{x})}{Q(\mathbf{x})} \right)$$

- ▶ 非負
- ▶ 三角不等式を満たさない、非対称

- ▶ データの経験分布とモデルの分布を合わせる.

- ▶ KL情報量の最小化

$$P_D(\mathbf{x}) = \frac{1}{D} \sum_{d=1}^D \delta(\mathbf{x} - \mathbf{x}^{(d)})$$

$$D_{\text{KL}}(P_D|P_{\mathbf{u}}) = \int d\mathbf{x} P_D(\mathbf{x}) \log \left(\frac{P_D(\mathbf{x})}{P(\mathbf{x}|\mathbf{u})} \right)$$



ボルツマン機械学習の基礎

▶ 最尤法

- ▶ データに対するモデルの尤もらしさ（対数尤度関数の経験平均）

$$L(\mathbf{u}) = \frac{1}{D} \sum_{d=1}^D \log P(\mathbf{x} = \mathbf{x}^{(d)} | \mathbf{u})$$

▶ 最尤法

- ▶ データに対するモデルの尤もらしさ（対数尤度関数の経験平均）

$$L(\mathbf{u}) = \frac{1}{D} \sum_{d=1}^D \log P(\mathbf{x} = \mathbf{x}^{(d)} | \mathbf{u})$$

- ▶ 統計力学のエントロピー最大原理と等価

$$L(\mathbf{u}) = -\frac{1}{D} \sum_{d=1}^D E(\mathbf{x} = \mathbf{x}^{(d)} | \mathbf{u}) - \log Z(\mathbf{u})$$



ボルツマン機械学習の基礎

- ▶ 単純な勾配法の適用
 - ▶ 勾配方向に更新して最大値を求める

$$\mathbf{u}[t + 1] = \mathbf{u}[t] + \eta \frac{\partial L(\mathbf{u})}{\partial \mathbf{u}}$$

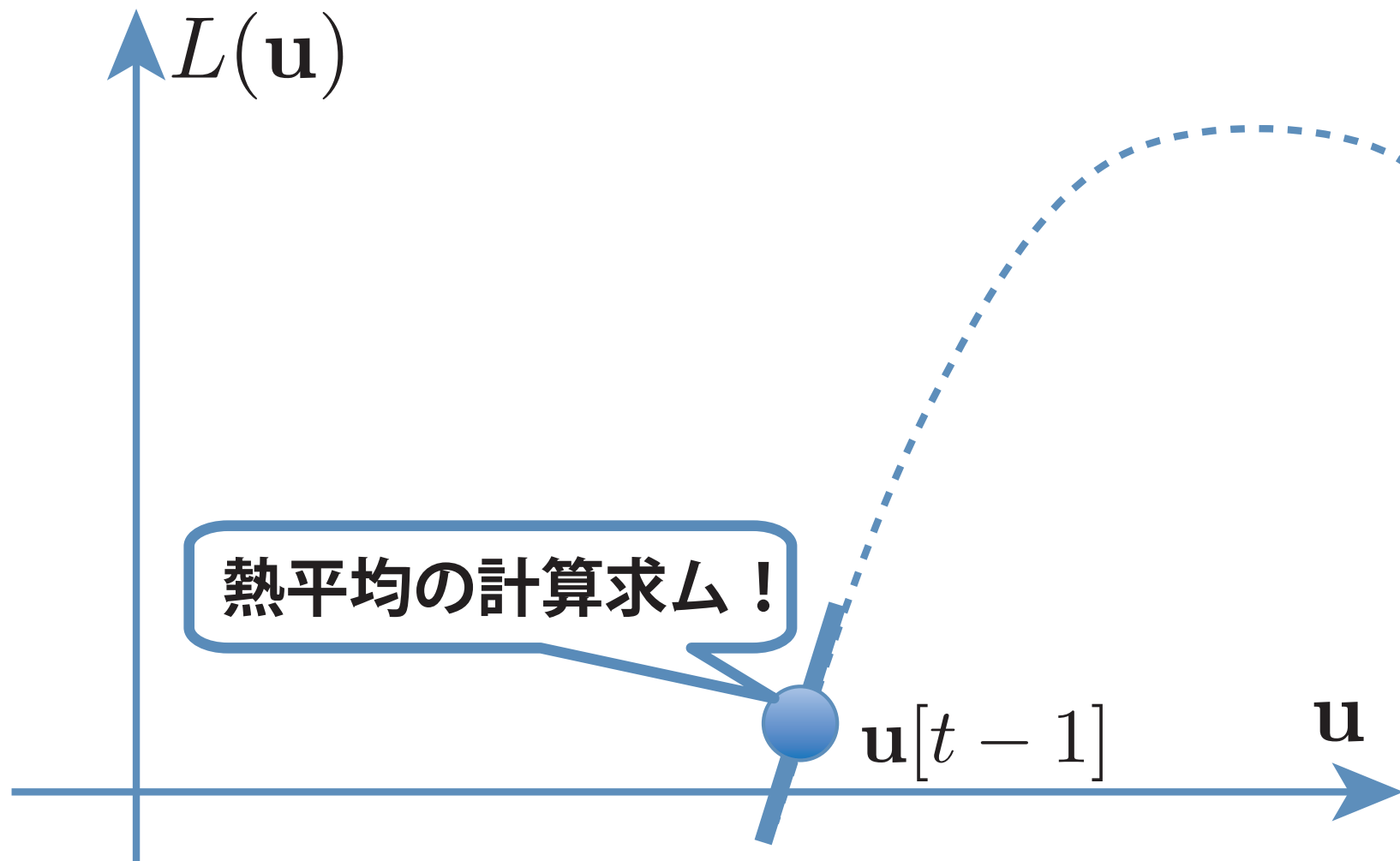
- ▶ η : 学習係数、更新幅を設定している

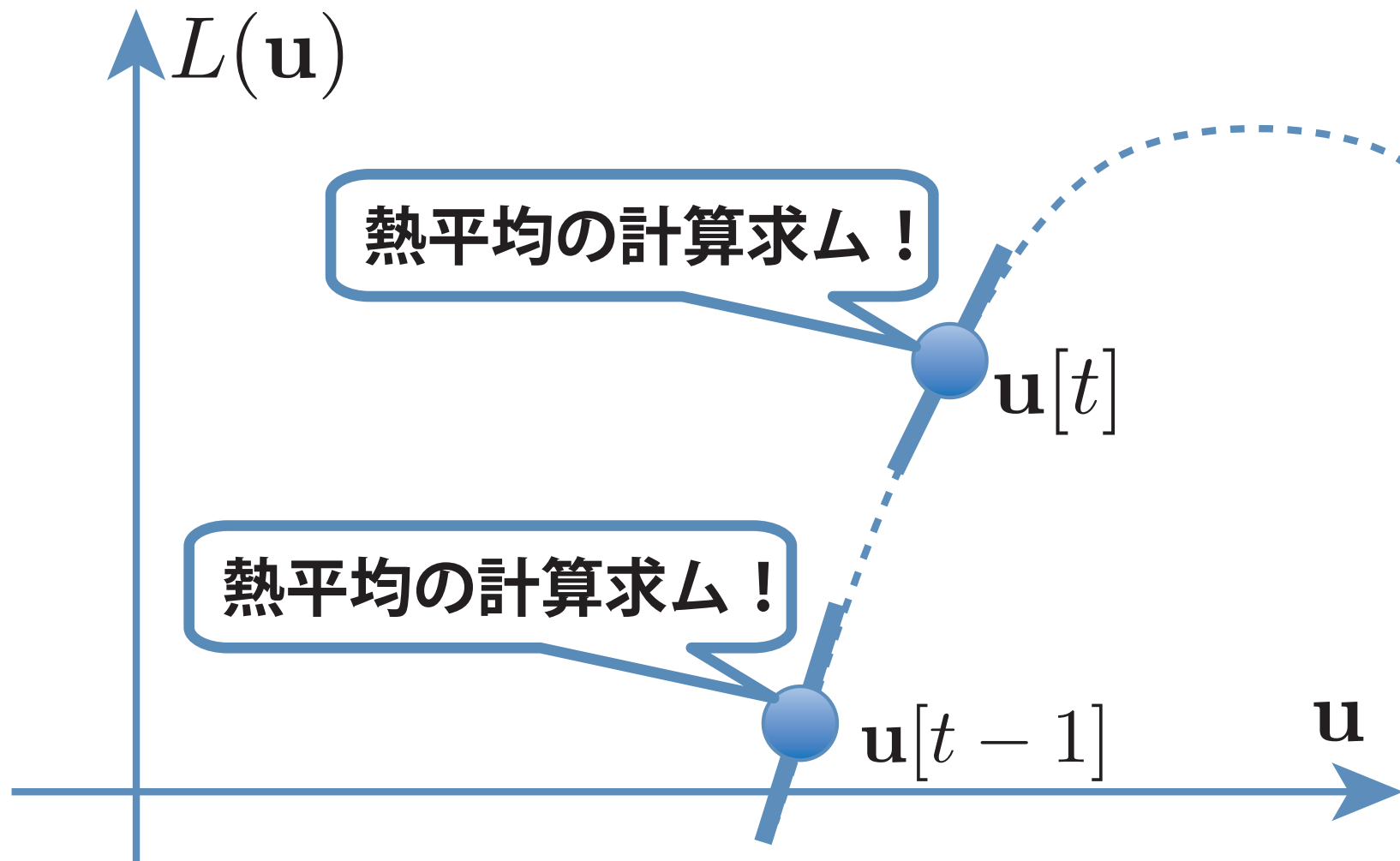
- ▶ 単純な勾配法の適用
 - ▶ 勾配方向に更新して最大値を求める

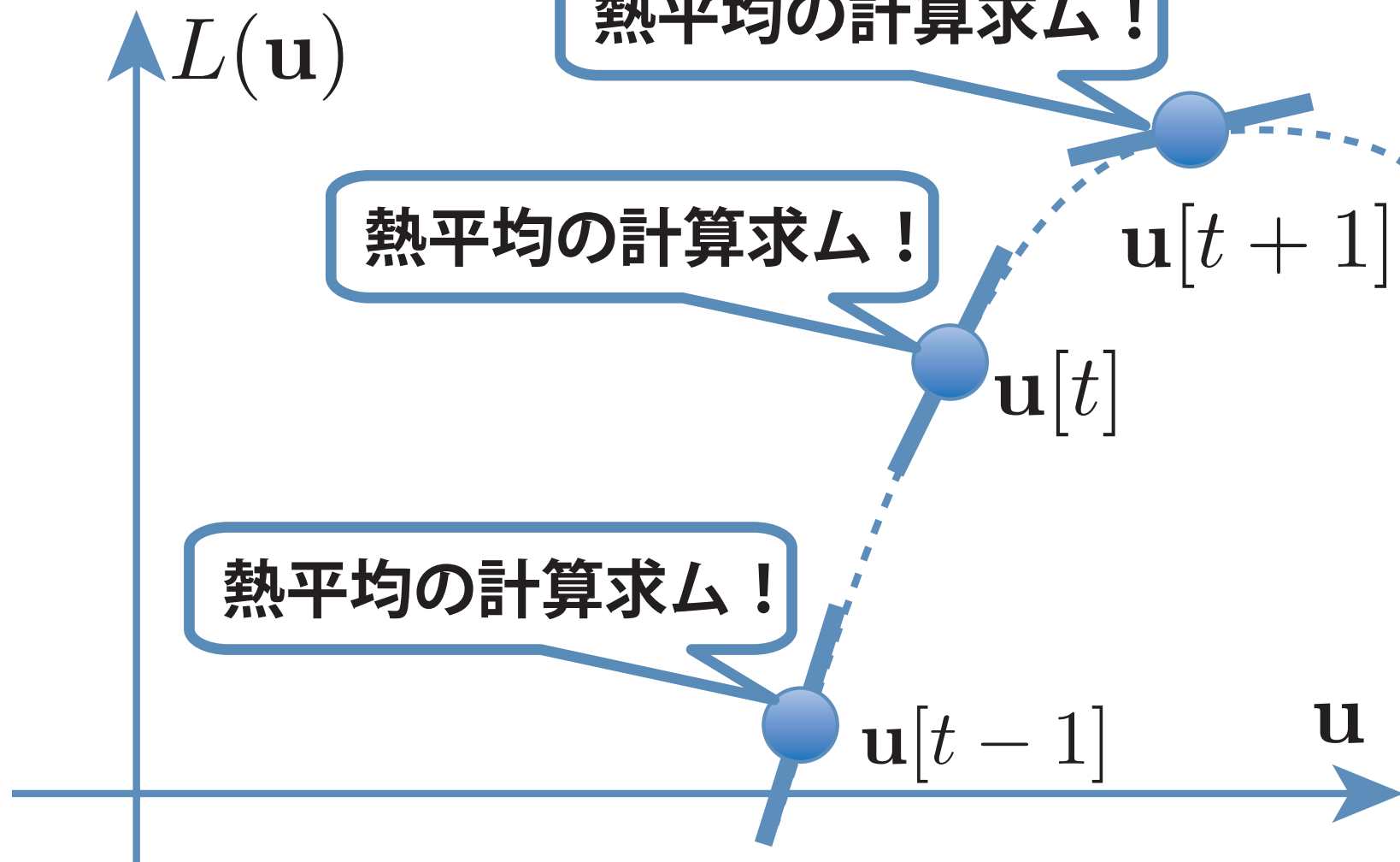
$$\mathbf{u}[t + 1] = \mathbf{u}[t] + \eta \frac{\partial L(\mathbf{u})}{\partial \mathbf{u}}$$

- ▶ η : 学習係数、更新幅を設定している
- ▶ じゃあ勾配を計算してみよう！

$$\frac{\partial L(\mathbf{u})}{\partial \mathbf{u}} = -\frac{1}{D} \sum_{d=1}^D \frac{\partial E(\mathbf{x} = \mathbf{x}^{(d)} | \mathbf{u})}{\partial \mathbf{u}} + \left\langle \frac{\partial E(\mathbf{x} | \mathbf{u})}{\partial \mathbf{u}} \right\rangle_{\mathbf{u}}$$







$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial \mathbf{u}} \right\rangle_{\mathbf{u}}$$

ボルツマン機械学習の危機

$$\log Z(\mathbf{u})$$

- ▶ イジング模型の場合
 - ▶ モーメントマッチング

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial J_{ij}} \right\rangle_{\mathbf{u}} = \langle x_i x_j \rangle_{\mathbf{u}}$$

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial h_i} \right\rangle_{\mathbf{u}} = \langle x_i \rangle_{\mathbf{u}}$$

- ▶ イジング模型の場合
 - ▶ モーメントマッチング

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial J_{ij}} \right\rangle_{\mathbf{u}} = \langle x_i x_j \rangle_{\mathbf{u}}$$

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial h_i} \right\rangle_{\mathbf{u}} = \langle x_i \rangle_{\mathbf{u}}$$

- ▶ 相関関数、局所磁化の計算が必要！
 - ▶ 統計力学??

$$m = \frac{1}{N} \sum_{i=1}^N x_i$$

- ▶ 期待値の計算はなぜ難しい？
 - ▶ クロスターム、相関の存在

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N \sum_{j \in \partial i} J_{ij} x_i x_j + \sum_{i=1}^N h_i x_i$$

▶ 期待値の計算はなぜ難しい？

- ▶ クロスターム、相関の存在

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N \sum_{j \in \partial i} J_{ij} x_i x_j + \sum_{i=1}^N h_i x_i$$

- ▶ 全結合ではない
- ▶ レプリカ法？
- ▶ Jとhで平均的な性質が知りたいわけでもない

▶ 独立な分布で近似しよう！

- ▶ 平均場近似

$$P(\mathbf{x}|\mathbf{u}) \approx \prod_{i=1}^N P_i(x_i|\mathbf{u})$$



- ▶ 独立って何も考えないで？

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N h_i x_i ??$$

- ▶ 独立って何も考えないで？

$$E(\mathbf{x}|\mathbf{u}) = \sum_{i=1}^N h_i x_i ??$$

- ▶ できるだけ元の問題に近く…

$$P_i^{\text{MF}}(x_i|\mathbf{u}) = \arg \min_{P_i(x_i|\mathbf{u})} \left\{ \int d\mathbf{x} \prod_{i=1}^N P_i(x_i|\mathbf{u}) \log \left(\frac{\prod_{i=1}^N P_i(x_i|\mathbf{u})}{P(\mathbf{x}|\mathbf{u})} \right) \right\}$$

- ▶ KL情報量の意味で尤もらしく
- ▶ 独立な確率分布だけど相関を取り込む

$$\text{s.t.} \quad \int dx_i P_i(x_i|\mathbf{u}) = 1$$

▶ ラグランジュ未定乗数法の適用

$$P_i^{\text{MF}}(x_i|\mathbf{u}) = \arg \min_{P_i(x_i|\mathbf{u})} \left\{ \int d\mathbf{x} \prod_{i=1}^N P_i(x_i|\mathbf{u}) \log \left(\frac{\prod_{i=1}^N P_i(x_i|\mathbf{u})}{P(\mathbf{x}|\mathbf{u})} \right) + \sum_{i=1}^N \lambda_i P_i(x_i|\mathbf{u}) \right\}$$

▶ ラグランジュ未定乗数法の適用

$$P_i^{\text{MF}}(x_i|\mathbf{u}) = \arg \min_{P_i(x_i|\mathbf{u})} \left\{ \int d\mathbf{x} \prod_{i=1}^N P_i(x_i|\mathbf{u}) \log \left(\frac{\prod_{i=1}^N P_i(x_i|\mathbf{u})}{P(\mathbf{x}|\mathbf{u})} \right) + \sum_{i=1}^N \lambda_i P_i(x_i|\mathbf{u}) \right\}$$

▶ 変分で確率分布をもとめる

$$P_i(x_i|\mathbf{u}) = \frac{1}{Z_i^{\text{MF}}(\mathbf{u})} \exp \left(-E_i(x_i|\mathbf{m}_{-/i}, \mathbf{u}) \right)$$

▶ ラグランジュ未定乗数法の適用

$$P_i^{\text{MF}}(x_i|\mathbf{u}) = \arg \min_{P_i(x_i|\mathbf{u})} \left\{ \int d\mathbf{x} \prod_{i=1}^N P_i(x_i|\mathbf{u}) \log \left(\frac{\prod_{i=1}^N P_i(x_i|\mathbf{u})}{P(\mathbf{x}|\mathbf{u})} \right) + \sum_{i=1}^N \lambda_i P_i(x_i|\mathbf{u}) \right\}$$

▶ 変分で確率分布をもとめる

$$P_i(x_i|\mathbf{u}) = \frac{1}{Z_i^{\text{MF}}(\mathbf{u})} \exp \left(-E_i(x_i|\mathbf{m}_{-/i}, \mathbf{u}) \right)$$

▶ イジング模型の場合

$$P_i(x_i|\mathbf{u}) = \frac{\exp \left(-\sum_{j \in \partial i} J_{ij} x_i m_j + h_i x_i \right)}{2 \cosh \left(\sum_{j \in \partial i} J_{ij} m_j + h_i \right)}$$



平均場近似

- ▶ 確率分布の一体近似の完成

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial h_i} \right\rangle_{\mathbf{u}} = \langle x_i \rangle_{\mathbf{u}}$$

- ▶ 確率分布の一体近似の完成

- ▶ 局所磁化についての自己無撞着方程式

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial h_i} \right\rangle_{\mathbf{u}} = \langle x_i \rangle_{\mathbf{u}}$$

$$m_i = \tanh \left(\sum_{j \in \partial i} J_{ij} m_j + h_i \right)$$

- ▶ 確率分布の一体近似の完成

- ▶ 局所磁化についての自己無撞着方程式

$$\left\langle \frac{\partial E(\mathbf{x}|\mathbf{u})}{\partial h_i} \right\rangle_{\mathbf{u}} = \langle x_i \rangle_{\mathbf{u}}$$

$$m_i = \tanh \left(\sum_{j \in \partial i} J_{ij} m_j + h_i \right)$$

- ▶ 磁化についての自己無撞着方程式：全体の一様な性質

$$m = \tanh \left(\frac{J}{T} m \right)$$

平均場近似を超えて

局所的な情報を推定できる計算手法

平均場近似

$$P(\mathbf{x}|\mathbf{u}) \approx \prod_{i=1}^N P_i(x_i|\mathbf{u})$$

局所磁化の自己無撞着方程式

$$m_i = \tanh \left(\sum_{j \in \partial i} J_{ij} m_j + h_i \right)$$

TAP方程式 (スピングラス理論由来)

平均場近似の改善

$$m_i = \tanh \left(\sum_{j \in \partial i} J_{ij} m_j + h_i + \sum_{j \in \partial i} J_{ij}^2 (1 - m_j)^2 m_i \right)$$

信念伝搬法

$$P(\mathbf{x}|\mathbf{u}) \approx \prod_{i,j} \frac{P_{ij}(x_i, x_j|\mathbf{u})}{P_i(x_i|\mathbf{u}) P_j(x_j|\mathbf{u})}$$

▶ 信念伝搬法

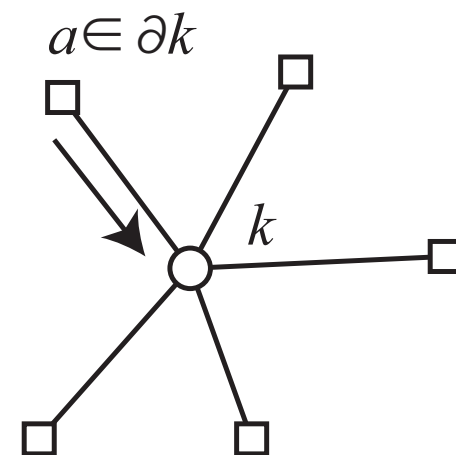
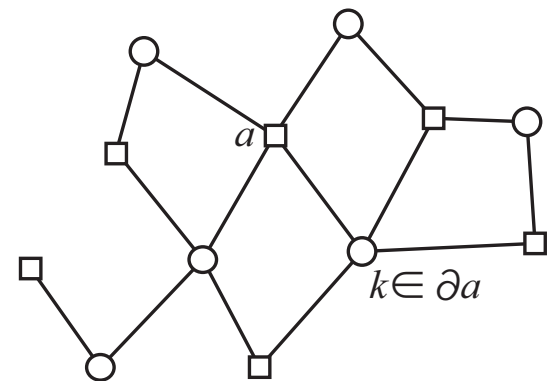
- ▶ 2体以上の相互作用もある一般的な場合

$$P(\mathbf{x}|\mathbf{u}) = \prod_k f_k(x_k) \prod_a f_a(\mathbf{x}_{\partial a})$$

- ▶ 得られる近似された確率分布

$$P(x_k) \propto f_k(x_k) \prod_{a \in \partial k} \tilde{M}_{a \rightarrow k}(x_k)$$

$$P(\mathbf{x}_{\partial a}) \propto f_a(\mathbf{x}_{\partial a}) \prod_{k \in \partial a} M_{k \rightarrow a}(x_k)$$



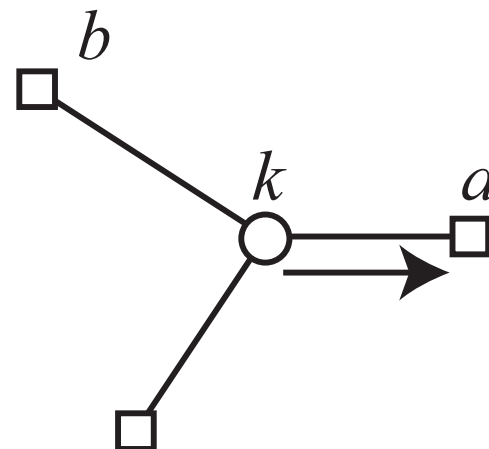
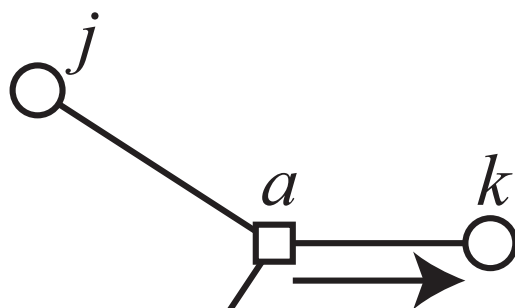
平均場近似を超えて

▶ 信念伝搬法

▶ メッセージ更新式

$$M_{a \rightarrow k}(x_k) \propto \int d\mathbf{x}_{\partial a / k} f_a(\mathbf{x}_{\partial a}) \prod_{j \in \partial a / k} \tilde{M}_{j \rightarrow a}(x_j)$$

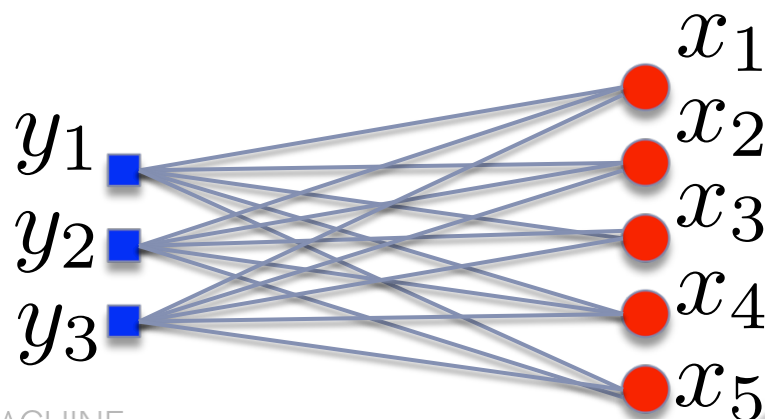
$$\tilde{M}_{k \rightarrow a}(x_k) \propto f_k(x_k) \prod_{b \in \partial k / a} M_{b \rightarrow k}(x_b)$$



- ▶ 圧縮センシングでも利用されている

- ▶ 基底追跡型最適化問題

$$P(\mathbf{x}|A, \mathbf{y}) \propto \prod_k \exp(-\beta|x_k|) \prod_{\mu} \exp\left(-\frac{1}{2\lambda}(y_{\mu} - \mathbf{a}_{\mu}^T \mathbf{x})^2\right)$$



- ▶ 圧縮センシングでも利用されている

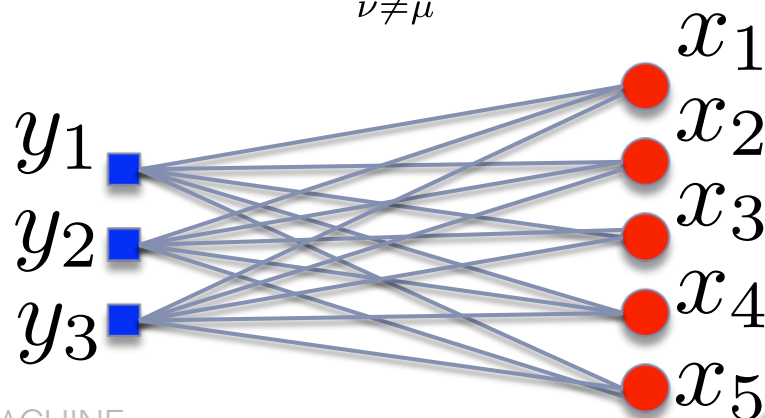
- ▶ 基底追跡型最適化問題

$$P(\mathbf{x}|A, \mathbf{y}) \propto \prod_k \exp(-\beta|x_k|) \prod_{\mu} \exp\left(-\frac{1}{2\lambda}(y_{\mu} - \mathbf{a}_{\mu}^T \mathbf{x})^2\right)$$

- ▶ メッセージ伝搬法

$$M_{\mu \rightarrow k}(x_k) \propto \int d\mathbf{x}_{/k} \exp\left(-\frac{1}{2\lambda}(y_{\mu} - \mathbf{a}_{\mu}^T \mathbf{x})^2\right) \prod_{l \neq k} \tilde{M}_{l \rightarrow \mu}(x_l)$$

$$\tilde{M}_{k \rightarrow \mu}(x_k) \propto \exp(-\beta|x_k|) \prod_{\nu \neq \mu} M_{\nu \rightarrow k}(x_k)$$



- ▶ 圧縮センシングでも利用されている

- ▶ 基底追跡型最適化問題

$$P(\mathbf{x}|A, \mathbf{y}) \propto \prod_k \exp(-\beta|x_k|) \prod_{\mu} \exp\left(-\frac{1}{2\lambda}(y_{\mu} - \mathbf{a}_{\mu}^T \mathbf{x})^2\right)$$

- ▶ メッセージ伝搬法

$$M_{\mu \rightarrow k}(x_k) \propto \int d\mathbf{x}_{/k} \exp\left(-\frac{1}{2\lambda}(y_{\mu} - \mathbf{a}_{\mu}^T \mathbf{x})^2\right) \prod_{l \neq k} \tilde{M}_{l \rightarrow \mu}(x_l)$$

$$\tilde{M}_{k \rightarrow \mu}(x_k) \propto \exp(-\beta|x_k|) \prod_{\nu \neq \mu} M_{\nu \rightarrow k}(x_k)$$

- ▶ 更に高速なメッセージ伝搬法を導出可能 (Approximate Message Passing)

$$\mathbf{v}[t] = \mathbf{x}[t] + \frac{A^T}{L\lambda}(\mathbf{y} - A\mathbf{x}) + \frac{1}{M} \sum_{k=1}^N S'_{1/L}(v_k[t-1]) A^T(\mathbf{y} - A\mathbf{x}[t-1])$$

$$\mathbf{x}[t+1] = S_{1/L}(\mathbf{v}[t])$$