学習理論の練習　10

学習理論の基礎

学習理論の数学的基礎

真の分布 $q(x)$, 学習モデル $p(x|w)$, 事前分布 $\varphi(w)$ が与えられたとする。

$$L(w) = -\int q(x) \log p(x|w)\, dx.$$

の最小値を与えるパラメータを w_0 とし、関数 $K(w)$ を

$$K(w) = \int q(x) \log \{ p(x|w_0)/p(x|w) \}\, dx$$

と定義する。学習理論のゼータ関数を

$$\zeta(z) = \int K(w)^z \varphi(w)\, dw \qquad (z \text{ in } \mathbf{C})$$

と定義すると、この関数は $\mathrm{Re}(z)>0$で正則であり、全複素平面に有理型関数として一意に解析接続できることを証明できる(Atiyah)。その極はすべて負の有理数である(Kashiwara)。原点に一番近い極を $-\lambda$ としその位数を m とするとき λを実対数閾値といい、m を多重度という。

学習の漸近理論

データの数（サンプルサイズ）n が無限大に近づくとき、汎化損失と自由エネルギーの平均値 $E[G_n]$ と $E[F_n]$ は次の漸近挙動を持つことが証明できる。

$$E[G_n] = L(w_0) + \lambda/n + o(1/n),$$

$$E[F_n] = n\,L(w_0) + \lambda \log n - (m-1)\log\log n + O_n,$$

ここで $o(1/n)$ は $(1/n)$ よりも早く零に近づく数列であり, O_n は定数に収束する数列である。（注: $E[G_n] = E[F_{n+1}] - E[F_n]$ が成り立っていることに注意）。

与えられた三組 $(q(x), p(x|w), \varphi(w))$ に対して実対数閾値を求めることは数学的な研究の課題である。

任意の解析関数 $w=g(u)$ により 三組$(q(x),\ p(x|g(u)),\ \varphi(w)|g'(u)|\ )$ に移行しても実対数閾値は変わらない（双有理不変）であることを利用して、探す方法が有効である。関数 $g(u)$ の探し方は数学でよく研究されている。

行列を分解する (1)

縮小ランク回帰

$X \in \mathbf{R}^M$, $Y \in \mathbf{R}^N$

$B \in \mathbf{R}^{NH}$, $A \in \mathbf{R}^{HM}$

$B_0 \in \mathbf{R}^{NH_0}$, $A_0 \in \mathbf{R}^{H_0M}$

真: $Y = B_0A_0X$ + 雑音

モデル: $Y = BAX$ + 雑音

データ $\{X_i\}$ がある確率分布から独立に得られ、
データ $\{Y_i\}$ が$\{X_i\}$ のあるランクの線形変換+雑音により得られたとする。このとき線形変換のランクを知りたい。
これはニューラルネットでシグモイド関数の代わりに恒等写像を用いた場合と等価なモデルである。

青柳の定理

定理（青柳美輝,2005） 縮小ランク回帰で真のランクを r とする。

(1) $N+r<M+H, M+r<N+H, H+r<M+N$ かつ$M+H+N+r$ が偶数のとき
$m=1$ で $\lambda=\{2(H+r)(M+N)-(M-N)^2-(H+r)^2\}/8$.

(2) $N+r<M+H, M+r<N+H, H+r<M+N$ かつ$M+H+N+r$ が奇数のとき
$m=2$ で $\lambda=\{2(H+r)(M+N)-(M-N)^2-(H+r)^2+1\}/8$.

(3) $M+H<N+r$ のとき $m=1$ で $\lambda=\{HM-Hr+Nr\}/2$.

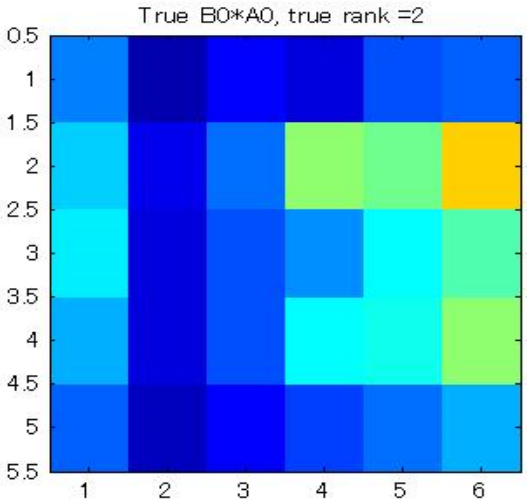
(4) $N+H<M+r$ のとき $m=1$ で $\lambda=\{HN-Hr+Mr\}/2$.

(5) $M+N<H+r$ のとき $m=1$ で $\lambda=MN2$.

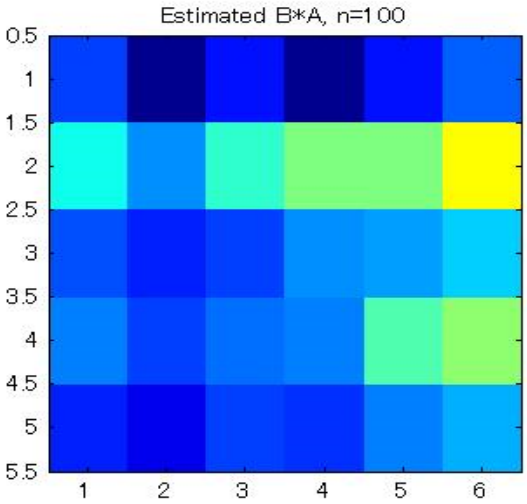
Drton and Plummer (2017, Royal Series B) では、この定理を用いて真の分布がわからないときにλを推測する統計的アルゴリズムを提案している。

具体例

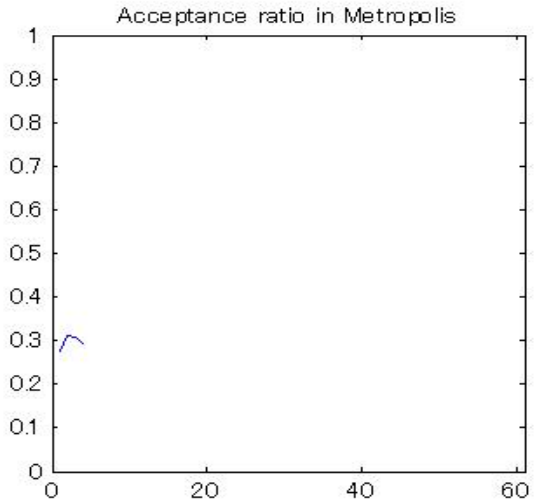
真の行列(5×6)
ランクは2



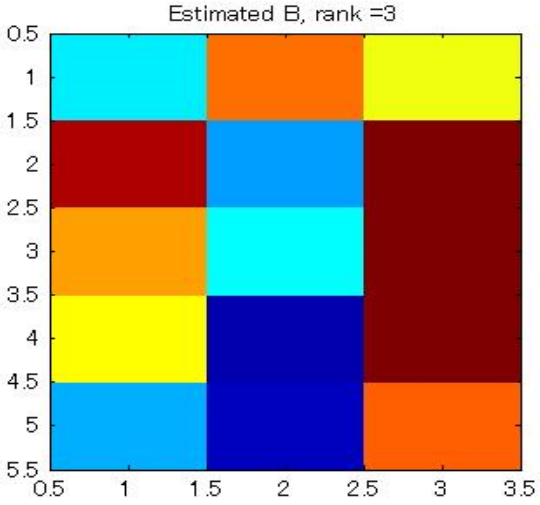
推測された行列(BA)



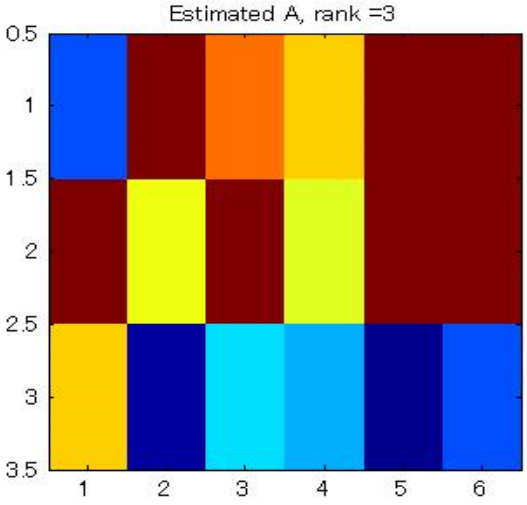
受容確率



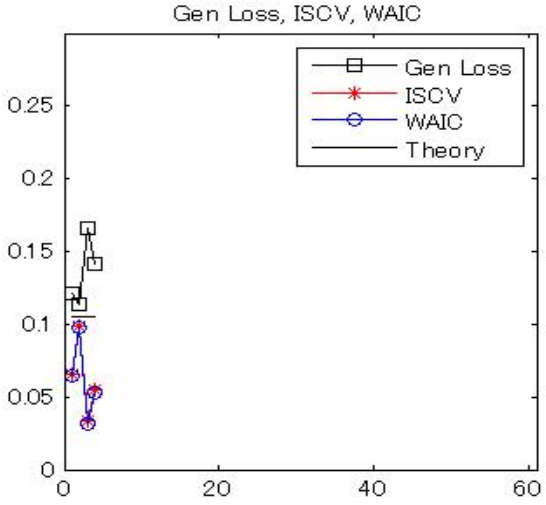
分解 B(5×3)



分解 A(3×6)



理論と実験



行列を分解する (2)

行列分解

$X \in \mathbf{R}^{MN}$

$C_0 \in \mathbf{R}^{MN}$, rank $C_0 = H_0$

$B \in \mathbf{R}^{NH}, A \in \mathbf{R}^{HM}$

真: $X = C_0 +$ 雑音

モデル: $X = BA +$ 雑音

データ $\{X_i\}$ がある行列 C_0 に雑音が加わって得られたとする。
このとき C_0 のランクを推定し、その行列分解を行いたい。
縮小ランク回帰とは統計学的な意味では異なるモデルであるが
同じ実対数閾値を持つことが知られている。

非負値行列分解

$X \in \mathbf{R}^{MN}$

$C_0 \in \mathbf{R}^{MN}$, rank $C_0 = H_0$

$B \in \mathbf{R}^{NH}, A \in \mathbf{R}^{HM}$, C_0, B,A の要素はすべて非負

真: $X = C_0$ + 雑音

モデル: $X = BA$ + 雑音

データ $\{X_i\}$ が要素がすべて非負の行列 C_0 に雑音が加わって得られたとする。このとき C_0 のランクを知り、非負の行列BAに分解をしたい。

非負値行列分解についての注意

◎ 画像や音声の信号処理や商品購買情報解析において、非負値による分解のほうが普通の分解よりも応用における有用性が見られたという研究報告があり（分解に差を使うことができないため）、多くの研究がある。

◎ 統計的推測の問題ではなく、単なる行列の分解 $C_0=BA$ で あっても、与えられた C_0 に対して B と A を見つけることは簡単でなく、適当な初期値から出発して繰り返し法で数値的に解くことが多い。有料の数学ソフトでも、計算するたびに結果が大きく異なることがある。

◎ 通常の行列のランクと非負値行列としてのランクは同じではない。

(例) $\mathrm{rank}\begin{pmatrix} 1 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 1 \end{pmatrix}$ =3 だが非負値ランクは4

線形代数の直感は成立たないことも多いので注意が必要です。

林の定理

定理（林直輝,2016）非負値行列分解で真の非負値ランクを r とする。

$$\lambda \leqq \{ (H-r)\min\{M,N\} + r(M+N-1)\}/2$$

なおλは縮小ランク回帰のとき以上の値である。

林(2017)ではさらにタイトなバウンドを導出している。

非負値行列のようにパラメータのとりうる範囲に制限があるときには、事後分布を作ることが簡単ではない（高次元であればあるほど難しくなる）。非負値行列分解の事後分布を精度よく作る方法も研究の課題である。

理論値がわかっているとMCMC法が良好に動作しているかどうかの確認にも利用可能である。

マルコフ連鎖モンテカルロ法

モンテカルロとは

モナコ公国の街の名前：カジノやF1で有名

モンテカルロ法とは
乱数を利用して数値計算を行う方法のこと

考えた人：スタニスワフ・ウラム（らしい）
名づけた人：フォン・ノイマン（らしい）

マルコフ連鎖とは

確率変数の数列 $\{X_n\}$ で X_n が $p(X_n|X_{n-1})$ によって生成される（1時刻前だけでなくもっと前からの影響を受ける場合も含む）。

「マルコフ連鎖」と言う場合には、上記の条件つき確率は時間的に変化しない場合が想定されていることが多い。（どんな確率変数もベイズの定理を使えば条件つき確率の連鎖でかける）。

なお、毎時間独立なものもマルコフ連鎖の特別な場合であるが、「マルコフ連鎖」は独立でない場合を意味していることが多い。

学習理論だけでなく多くの科学の領域で大切な役割を果たしているにも関わらず、大学での講義が少ないため社会に出てから困っている人が多いもののひとつである。

目標

関数 H(w) が与えられたとき確率密度関数

$$p(w)=(1/Z)\exp(-nH(w))$$

による平均値を計算したい。つまり積分値

$$\int f(w)\,p(w)\,dw$$

を求めたい。うまく確率変数 $w_1,w_2,\ldots,w_K$ を生成して

$$\int f(w)\,p(w)\,dw = \lim_{K\to\infty} (1/K)\sum_{k=1}^{K} f(w_k)$$

となるようにしたい。

（注）Z はこの方法だけでは計算できないのでさらに高度な方法が必要。

マルコフ過程

どうしたら $w_1,w_2,\ldots,w_K$ を生成できるか。

うまいマルコフ過程 $p(w_2|w_1)$ を使って生成したい。

次の二つが成り立つことが十分条件

(1) 集合 $\{w;p(w)>0\}$ の任意の点の近傍に到達する確率が0ではない。

(2) 詳細釣合い。任意の (w_1,w_2) について
$p(w_2|w_1)\ p(w_1) = p(w_1|w_2)\ p(w_2)$

（注）十分条件であるが必要条件ではない。順番を決めたギブスサンプラーは詳細釣り合いを満たさないが、前ページの目標に使うことができる。

いろいろなMCMC。

（１）メトロポリス法

（２）ギブスサンプラー法

（３）ハミルトニアンモンテカルロ法

（４）ランジュバン法

（５）マルチカノニカル法

（６）レプリカ交換法

メトロポリス法

メトロポリスは考案した人の名前です。

メトロポリス法

r(u|v)=r(v|u) を満たす条件つき確率(例：正規分布）を準備。

(1) w_1 を初期化。i=1とする。

(2) $r(w|w_i)$ から w をサンプリング。

(3) もし $\exp(-H(w)+H(w_i)) > \mathrm{rand}()$ なら

$w_{i+1}=w$.

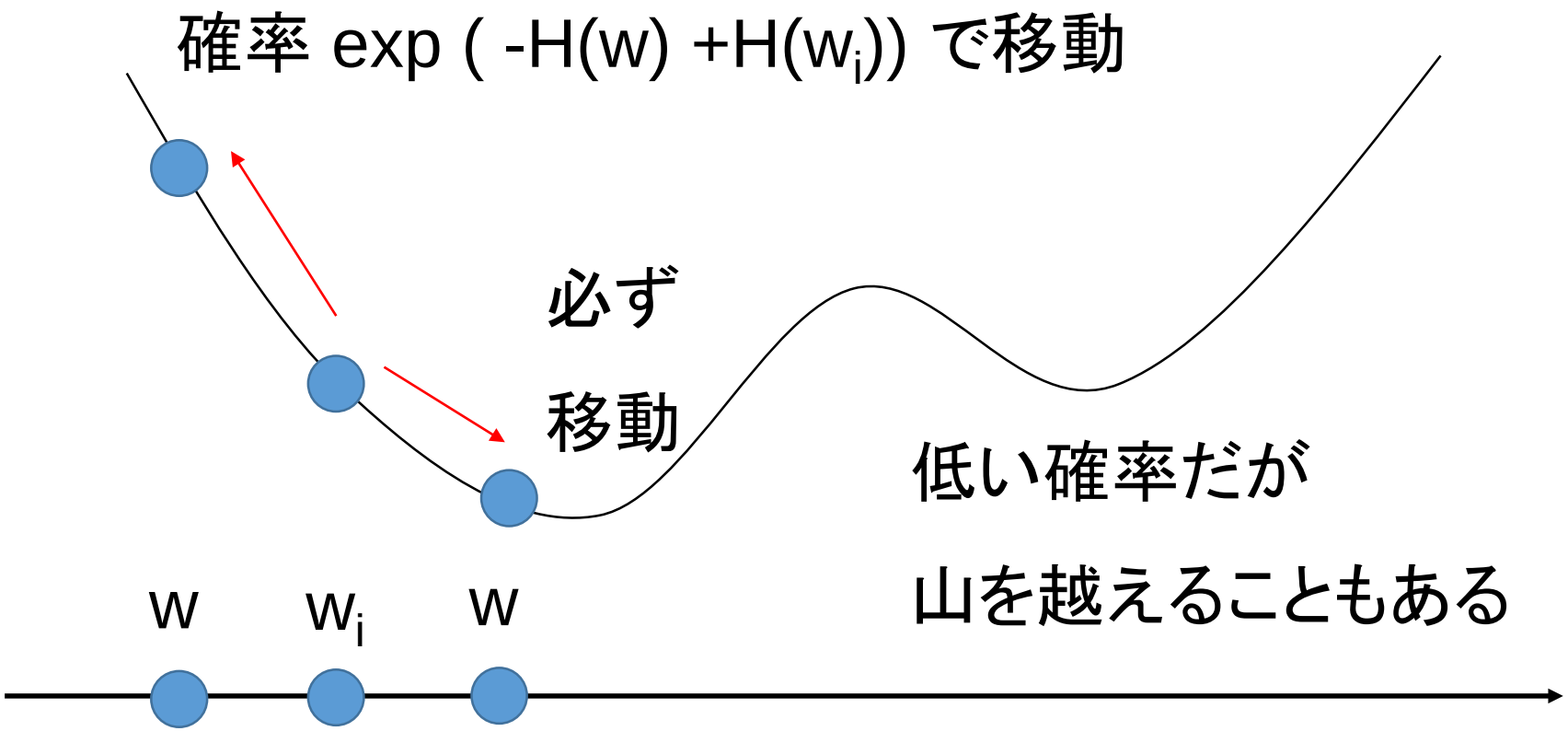
そうでないなら $w_{i+1}=w_i$ 。

[0,1]上の一様乱数

(4) i=i+1 として(2)に戻る。

このとき $p(w_{i+1}|w_i)$ は詳細釣合いを満たす。

メトロポリス法を納得する。



メトロポリス法の設計

（1）最初の点

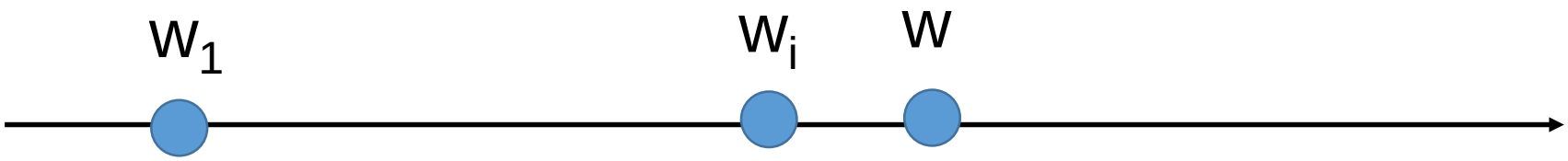
（2）山越え　谷渡り

（3）一歩の歩幅　受容確率を参考に

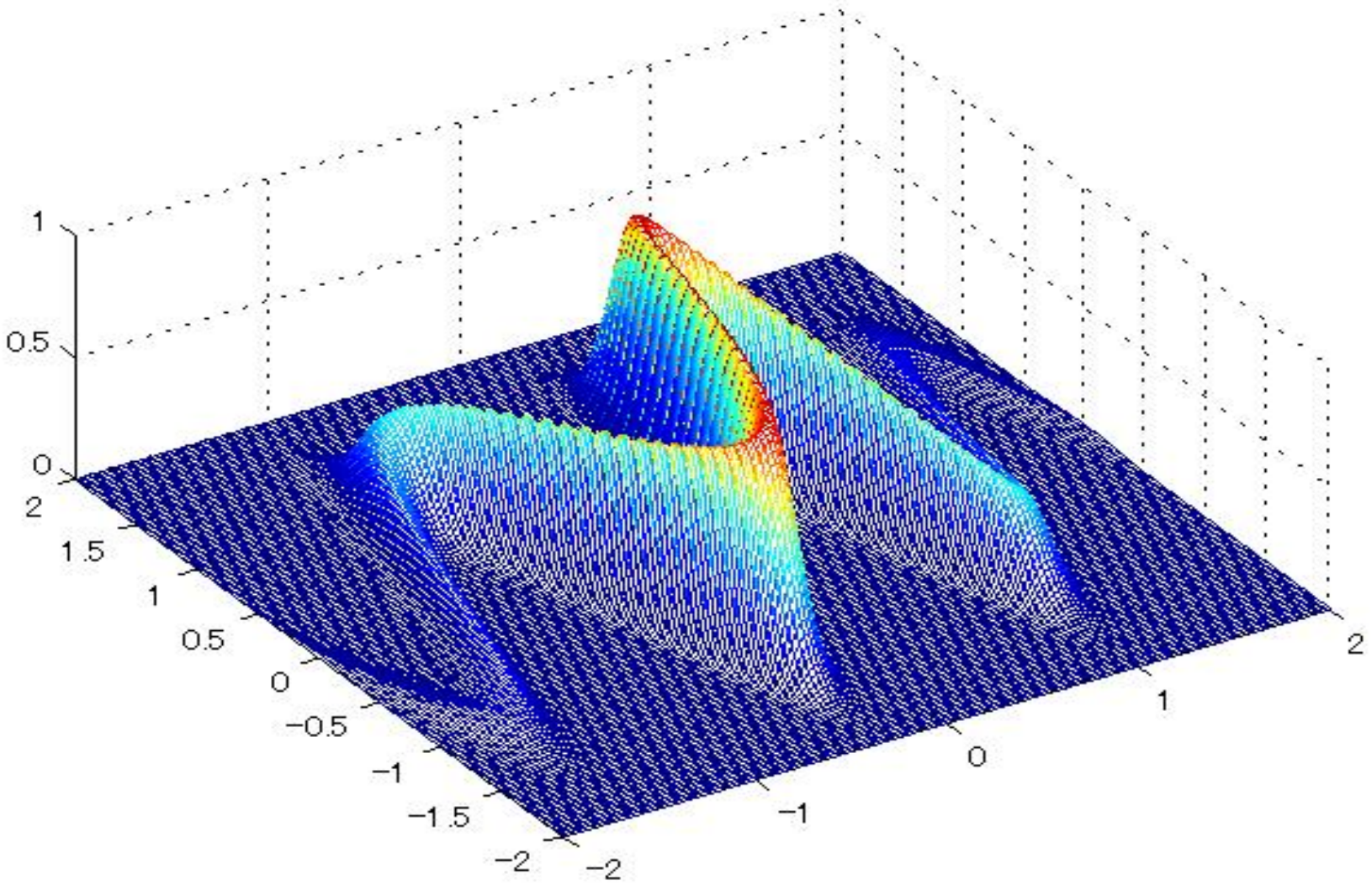
（4）バーンイン

（5）いくつ置きに

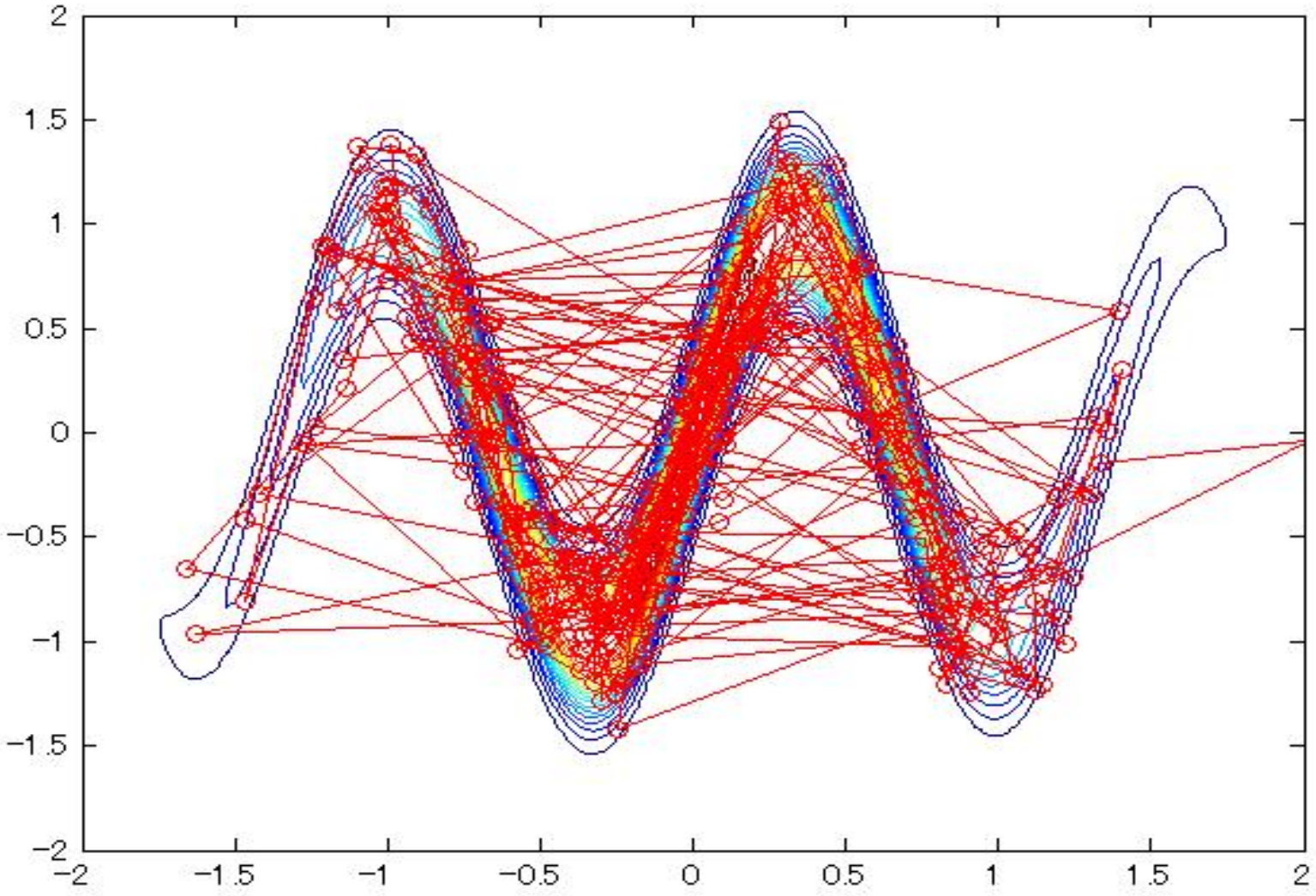
（6）どれだけたくさん



やってみる



やってみる



高次元空間

（1） 高次元空間の積分は確率的に実行するほうが有利である
といわれている。

(2） しかし高次元空間は目で見ることはできないので
MCMCが良好に動作しているかどうかを確認するのは難しい。
統計的に検定する方法はある(Geweke, Gelman-Rubinなど）が、
大規模な問題では検定を通過するのは容易でない（深層学習など）。

(3) 非負値行列分解のようにパラメータのとる範囲が限定されているときに良い工夫が作れると良い。

(4) Uターンなしハミルトン法が実装されたSTANは、ユーザーが任意に式で与えた分布に対してモンテカルロ・サンプリングを実行してくれるプログラム言語であり評価が高い。