

*Pattern
Recognition
and
Machine
Learning*

第11章 サンプルリング法

博士課程 1 年
原 祐輔

11章の内容

- 基本的なサンプリングアルゴリズム
 - 棄却サンプリング・適応的棄却サンプリング
 - 重点サンプリング
 - SIR
 - サンプリングとEMアルゴリズム
 - データ拡大アルゴリズム
- マルコフ連鎖モンテカルロ
 - Metropolis-Hastingsアルゴリズム
- ギブスサンプリング
- スライスサンプリング
- ハイブリッドモンテカルロアルゴリズム

本当の目次

- 疑似乱数の与太話
- not MCMC サンプリング
 - 逆関数法
 - 棄却サンプリング
 - 適応的棄却サンプリング
 - 重点サンプリング
 - SIR
- MCMCサンプリング
 - ギブスサンプリング
 - MH法

ありがちな話

- 「適当に乱数発生させて推定すればいいよね」
- 「そこは乱数でシミュレーション積分して...」
- 「あとは計算機がやってくれるから...」

- **だが、待ってほしい**
- 適切な分布からサンプリングするのは難しいという声が聞こえないだろうか
- 乱数によるサンプリングが単純ではないという声に謙虚に耳を傾けるべきではないだろうか

ということで

- PRML11章の内容を踏襲しつつ，順番をちょっと入れ替えたりしてサンプリングについて整理する
- 特にMCMCでないサンプリングとMCMC的サンプリングを理解することに重点
- 「適当に乱数発生させて～」ということがどれだけ難しいか，またそのサンプリングの使い方の難しさをきっちり体感する
- 理論的な展開は今回はある程度省く
- ギブスサンプラーやMetropolis-Hastingsがなぜあれほどに賞賛され，幅広い分野で使われているのかを理解することが目的

全体像を俯瞰

逆関数法

棄却サンプリング

適応的棄却サンプリング

重点サンプリング

SIR

データ拡大サンプリング

マルコフ連鎖モンテカルロ

メトロポリス・ヘイスティンクス法

ギブス・
サンプリング

メトロポリス法


スライスサンプリング

その前に乱数（疑似乱数）について

- 乱数を発生させるのはそれほど簡単か？
- とりあえず 1 ～ 10 の間で 10 個乱数を発生させてみてください
- たぶん無理

- 計算機はどのように乱数を発生させているのか？
- 合同乗算法...使ってはいけなと言われて続けてきた
- トーズワース法・シフトレジスター法...推奨されてきた
- メルセンヌ・ツイスター法...いまならこれ！

メルセンヌ・ツイスター法

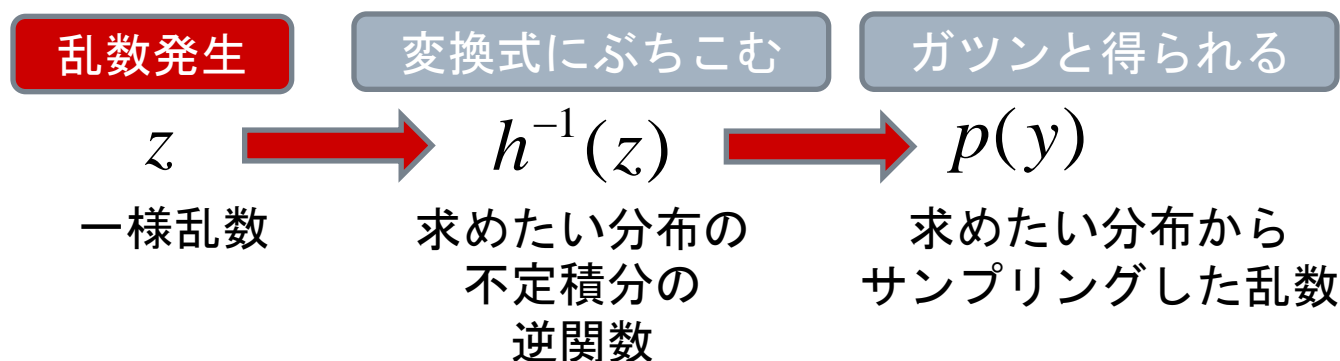
- 1996年から1998年に松本眞と西村拓士によって 開発された擬似乱数生成器
- $2^{19937}-1$ という長い周期
- あらゆる擬似乱数生成法の中でもっとも速い（当時）
- 開発当初は Primitive Twisted Generalized Feedback Shift Register Sequence という名前であったが、クヌースに名前が長すぎると言われたため現在の名前に変更された。
- Mersenne Twister の MT には、開発者の名前 「まこと」と「たくし」のイニシャルという意味もこめられている。
-  デフォルトの乱数がメルセンヌ・ツイスタ

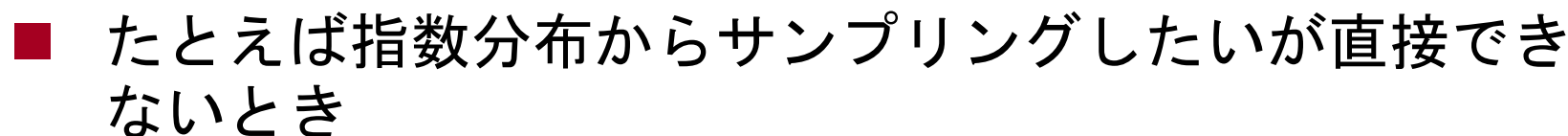
いまからサンプリング法の説明

- とにかく目的を見失わないこと
- 目的は「ある想定した分布」に従う乱数を発生させること（サンプリングすること）
- それが複雑で一般的な分布でもうまくいくように改善されていたり，サンプリング速度が速く効率的になるように改善されていったりしている
- とにかく目的は欲しい分布からサンプリングすること

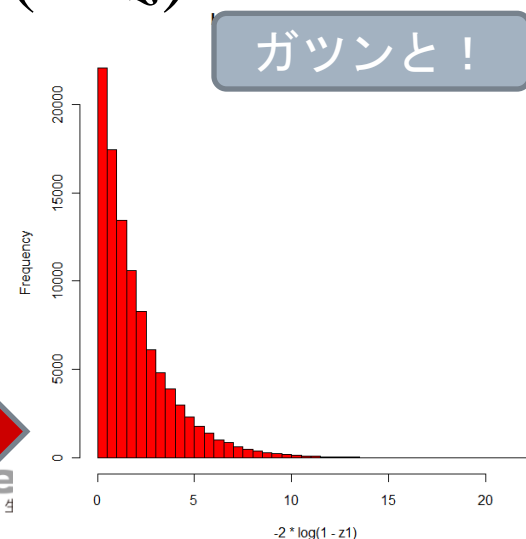
逆関数法

- まずは単純な逆関数法から
- 一様分布からサンプリングする（乱数を発生させる）ことができることが前提(MT法などで)
- 求めたい分布の確率密度関数の逆関数が簡単に書き下すことができる場合，次のフローでうまくいく．





よって、次の逆関数が得られる $h^{-1}(z) = -\lambda^{-1} \ln(1-z)$



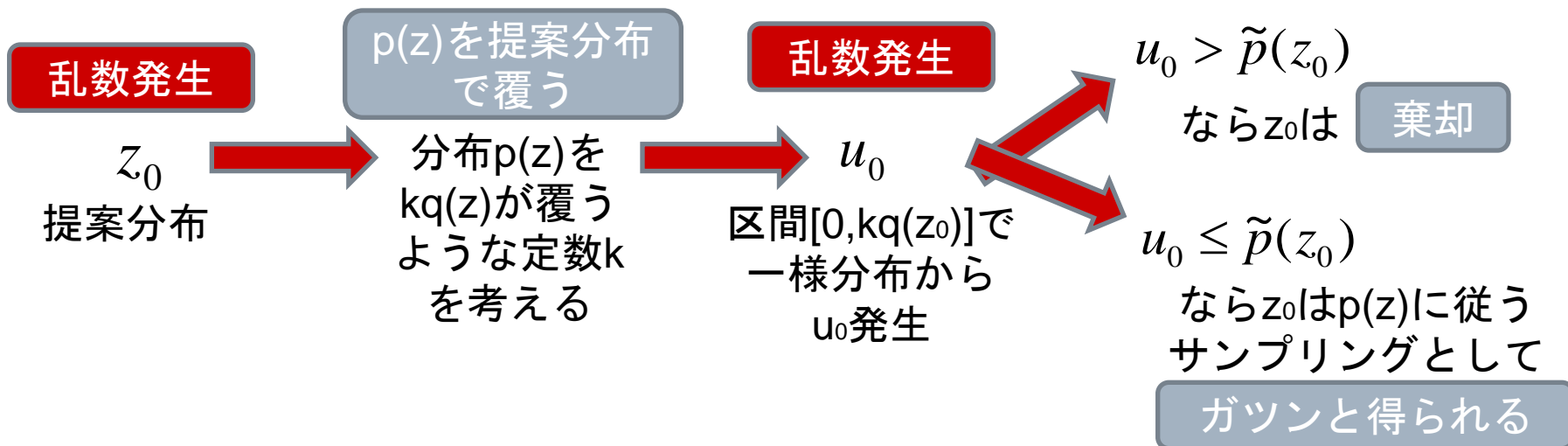
棄却サンプリング

- 逆関数が求められない**比較的複雑な分布**だってある
- やや複雑な分布 $p(z)$ からサンプリングしたいが、**直接** $p(z)$ から**サンプリングするのは困難**であるとする
- 任意の z が与えられたとき、**正規化定数 Z を除いた $p(z)$ を求めることは容易**であるとしよう（これはよくあること）

$$p(z) = \frac{1}{Z_p} \tilde{p}(z) \quad \tilde{p}(z) \text{はわかるが } Z_p \text{ わからん}$$

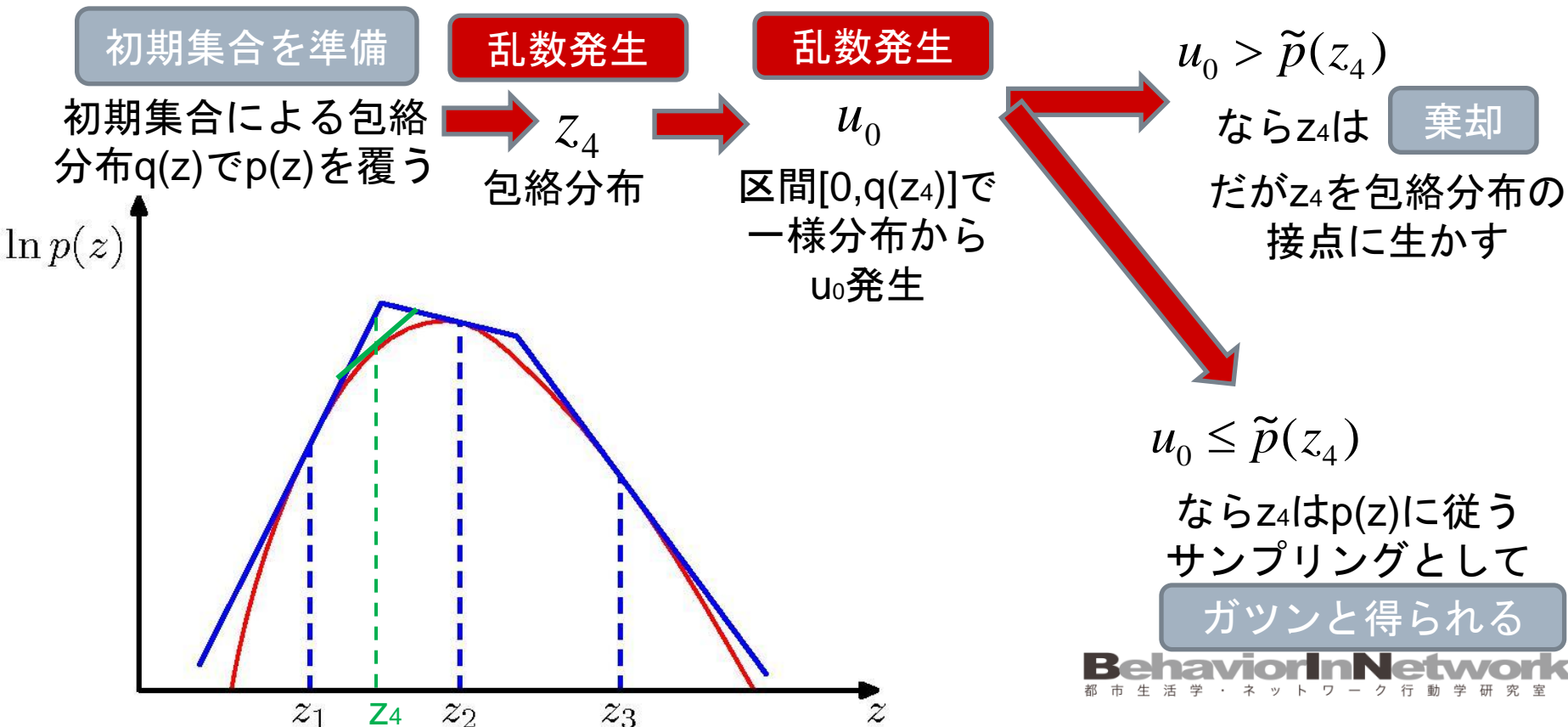
- ここで、容易にサンプリングできる**提案分布 $q(z)$** を考える
- これがミソ

棄却サンプリング



適応的棄却サンプリング

- 棄却サンプリングを適用したい多くの場合，適切な提案分布 $q(z)$ を解析的に決定することは難しい
- 別のアプローチとして，分布 $p(z)$ の観測値に基づいてその場で**包絡関数**を構築する方法



ここまでのまとめ

■ 逆関数法

- 解析的に正しいし，無駄がないというメリット
- 複雑な分布だと逆関数を求めるのはまず不可能というデメリット

■ 棄却サンプリング

- 比較的複雑な分布に対してもサンプリングできるメリット
- 適切な提案分布，定数 k を選ぶことが難しい
- 結果としてせっかくサンプリングしても大量に棄却するので無駄

■ 適応的棄却サンプリング

- 包絡関数を提案分布とすることで棄却が少なくなるメリット
- 棄却した値も新たな包絡線の接点として取り込む再活用
- 接線の計算負荷．また多次元で多峰性，するどいピークをもつ分布だとまず対応できないデメリット

重点サンプリング

- 分布 $p(z)$ からサンプリングしたいのではなくて、分布 $p(z)$ の**期待値を計算したいことが目的**のときもある
- しかし、 $p(z)$ から直接サンプリングするのは現実的ではないので、**棄却サンプリング的に提案分布を利用したい**

$$E[f] = \int f(z)p(z)dz = \int f(z) \frac{p(z)}{q(z)} q(z)dz$$

$$\approx \frac{1}{L} \sum_{l=1}^L \boxed{\frac{p(z^{(l)})}{q(z^{(l)})}} f(z^{(l)})$$

重要度重みと呼び、求めたいものとは異なった分布からサンプリングするバイアスを補正する。棄却サンプリングと異なり、すべてのサンプルを保持することに注意！

たとえば対数正規分布 $\text{LN}(1.1, 0.6)$ の平均を求めたいとき、 z を一様分布 $[0, 60]$ からサンプリングし、その値を用いて $p(z)$ を求める。また $q(z)=1/60$ であり、 $f(z)=z$ なので、 $p(z)*60*z$ によって Σ の中身が求まり、**期待値**が計算できる

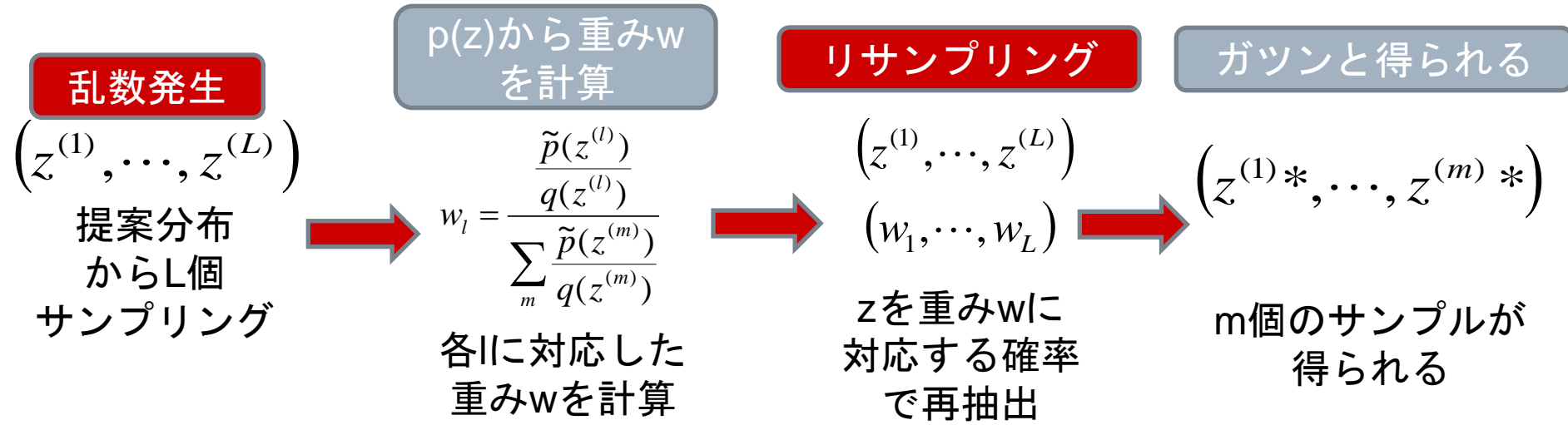
SIR(sampling-importance-resampling)

- 棄却サンプリングのうまい $q(z)$ と k の選び方を発見することとは現実的ではない
- そこで、重点サンプリングの重要度重みを有効活用してなんかうまいサンプリング方法はないか？→SIR
- 第1段階で提案分布から L 個のサンプル z を抽出する
- 第2段階で次式によって重み w を計算する

$$w_l = \frac{\frac{\tilde{p}(z^{(l)})}{q(z^{(l)})}}{\sum_m \frac{\tilde{p}(z^{(m)})}{q(z^{(m)})}}$$

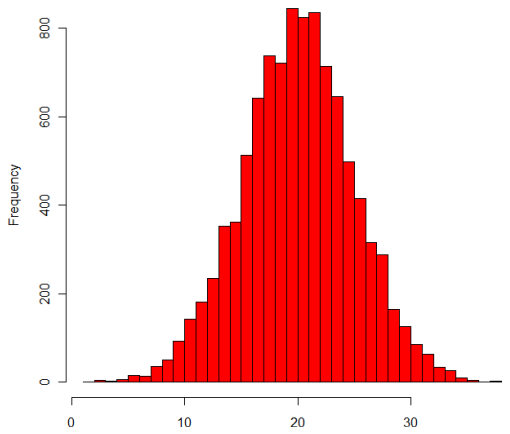
- 最後に z から重い w で与えられる確率に従ってリサンプリングする
- これは $L \rightarrow \infty$ では分布は正確に従うことが証明されている

SIR(sampling-importance-resampling)



ex) 提案分布 : $[0,50]$ の一樣分布で $p(z)$ が $N(20,5)$ とする

17.08	6.730e-02	1.654e-01	17.08
2.745	2.069e-04	5.087e-04	21.15
38.93	6.136e-05	1.508e-04	17.08
38.40	9.096e-05	2.235e-04	21.34
$z = 21.34$	$p(z) = 7.694e-02$	$w = 1.891e-01$	$z^* = 21.34$
41.04	1.133e-05	2.785e-05	21.15
20.52	7.935e-02	1.950e-01	20.52
22.17	7.259e-02	1.784e-01	22.17
21.15	7.770e-02	1.910e-01	21.15
13.30	3.254e-02	7.999e-02	17.08



ブートストラップ法
みたいな感じ？

サンプリングとEMアルゴリズム

- モンテカルロ法は**ベイズ的枠組み**だけでなく、最尤解を求める**頻度主義的**なパラダイムにおいても使える！
- 特にEMアルゴリズムにおいてEステップを解析的に実行できないモデルにおいて、サンプリング法がEステップを近似的に実行することが可能
- これを**モンテカルロEMアルゴリズム**と呼ぶ
- また、完全なベイズアプローチに移った**データ拡大アルゴリズム**と呼ばれる
Iステップ(imputation step) (Eステップに類似)と
Pステップ(posterior step) (Mステップに類似)を
交互に行うアルゴリズムも存在する
- こいつはほんとはMCMCの仲間

ここまでのまとめ（２）

■ 重点サンプリング

- 提案分布との重み付けで期待値を近似的に計算する
- 分布全体からサンプリングするものではない

■ SIR

- 棄却サンプリングと重点サンプリングの合わせ技
- あくまで最初のサンプリングからのリサンプリングなので、反復回数が少ないと偏る...が、 ∞ では近似できることは証明されてる

■ モンテカルロEMアルゴリズム

- サンプリングをEステップに代えて

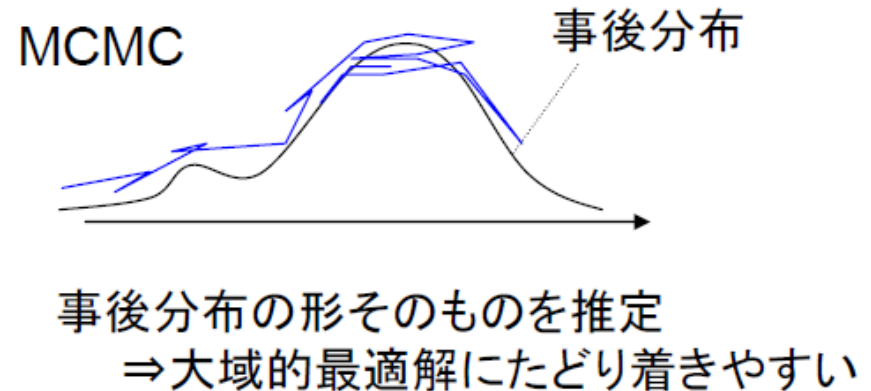
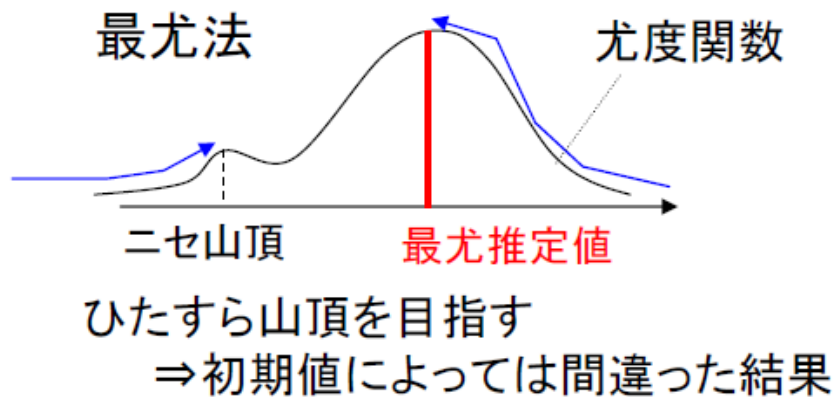
■ データ拡大法

- EMアルゴリズムのようにIステップとPステップでパラメータ θ を事後分布からサンプリング

MCMCの世界へ

頻度主義とベイズの世界観の違い

■ 最短経路の山登りかそれとも酔っぱらいの回遊か

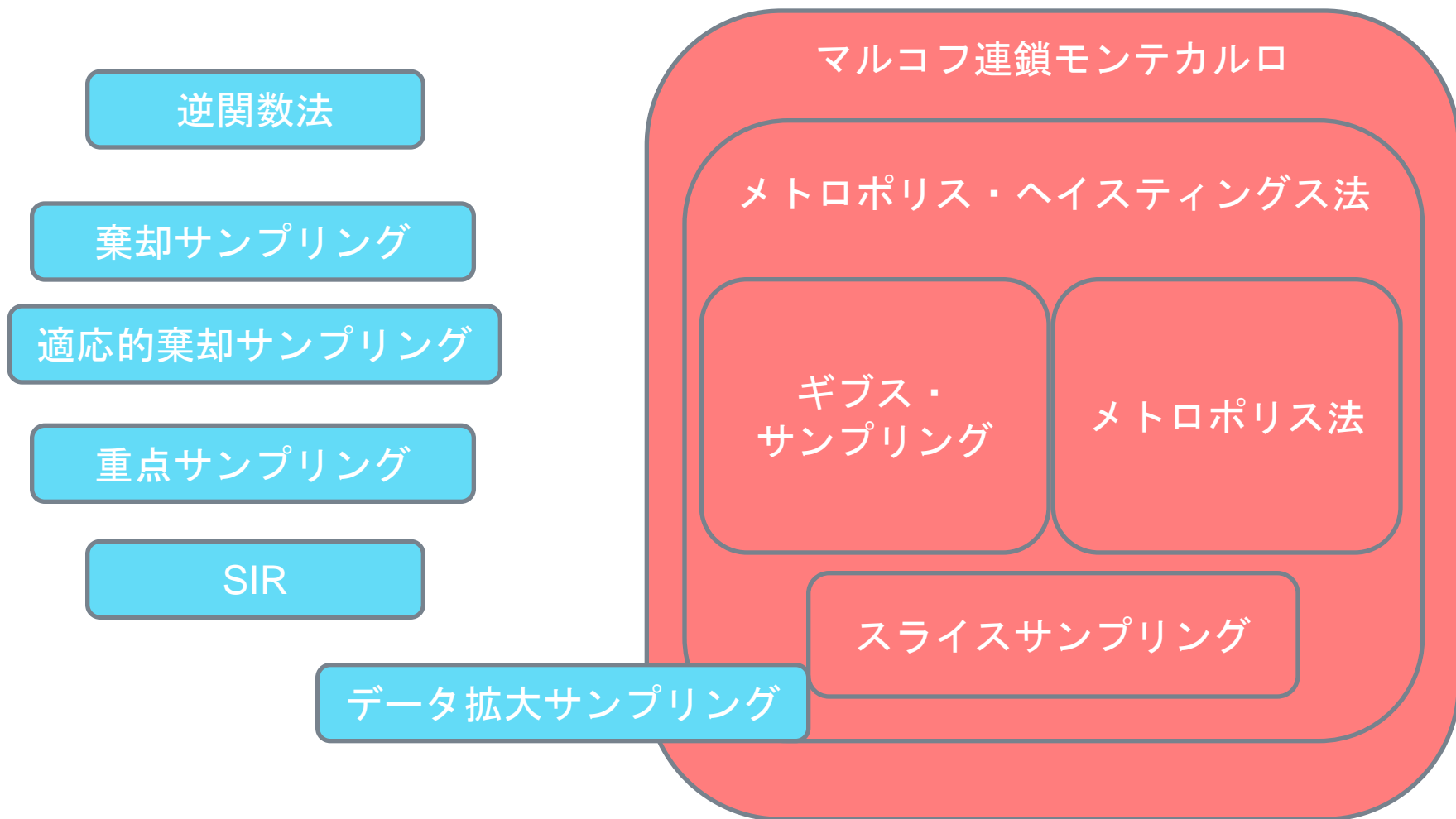


- 最尤法では山頂にのみ関心があるが
- MCMCは山全体の形に興味がある

なぜMCMCはモテ系なのか？

- 棄却サンプリングやSIR法では効率良いサンプリングのためにサンプラー（提案分布）を上手に選んでやる必要がある
- 非常に複雑な分布（非線形モデル）や大量のパラメータがある（次元が多い）とき、適切にサンプラーを選ぶのは難しい。
ときに不可能
- MCMCなら定常分布に収束するという性質を持っており、初期値に依存せず、いい感じのサンプリングが楽に行える
- モテ系というよりマッチョ系

全体像を俯瞰（再掲）



ギブス・サンプリング

- ギブス・サンプラーは
「条件付き分布からのサンプリング」
- サンプリングしたい \mathbf{x} を $\mathbf{x}=\{x_1, \dots, x_N\}$ と書こう． 毎回，
ひとつの成分 x_i を選び， その値を忘れて新しく取り直
すとする． その新しい x_i の値をそれ以外の成分を固定
した条件付き確率 $P(x_i | x_1, \dots, x_{i-1}, x_{i+1}, \dots, x_N)$ で選ぶ．
- これぞギブスサンプラーだ！

ギブス・サンプリング

1. $\{z_i : i = 1, \dots, M\}$ の初期値 $\{z_1^{(1)}, z_2^{(1)}, \dots, z_M^{(1)}\}$ を与える
2. $\tau = 1, \dots, T$ に対して以下を行う
 - $z_1^{(\tau+1)} \sim p(z_1 | z_2^{(\tau)}, \dots, z_M^{(\tau)})$ をサンプリングする
 - $z_2^{(\tau+1)} \sim p(z_2 | z_1^{(\tau+1)}, \dots, z_M^{(\tau)})$ をサンプリングする
 - ...
 - $z_j^{(\tau+1)} \sim p(z_j | z_1^{(\tau+1)}, \dots, z_M^{(\tau)})$ をサンプリングする
 - ...
 - $z_M^{(\tau+1)} \sim p(z_M | z_1^{(\tau+1)}, \dots, z_{M-1}^{(\tau+1)})$ をサンプリングする
3. 収束したと判定されるまでステップ2を繰り返す

一つ前の期の自分以外に依存しているのでマルコフ連鎖の一種である

ギブス・サンプリング

■ たとえば二変量正規分布の場合

(本当はもっと高次元がMCMCの腕の見せ所であるが...)

$$p(x_1, x_2) = \frac{1}{Z} \exp\left(-\frac{x_1^2 - 2bx_1x_2 + x_2^2}{2}\right)$$

ここで、一方を固定したときの条件付き密度は

$$p(x_1 | x_2) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_1 - bx_2)^2}{2}\right) \quad p(x_2 | x_1) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{(x_2 - bx_1)^2}{2}\right)$$

単なる正規分布からのサンプリングになったぞー！→余裕

- 条件付き分布からのサンプリングが簡単な場合, **魅力的**
- 条件付き分布が適切に書き下せる場合がギブスサンプラーの使いどころ
- 条件付き分布からのサンプリングが簡単でない場合→MH

ギブス・サンプリング

■ たとえば二変量正規分布の場合

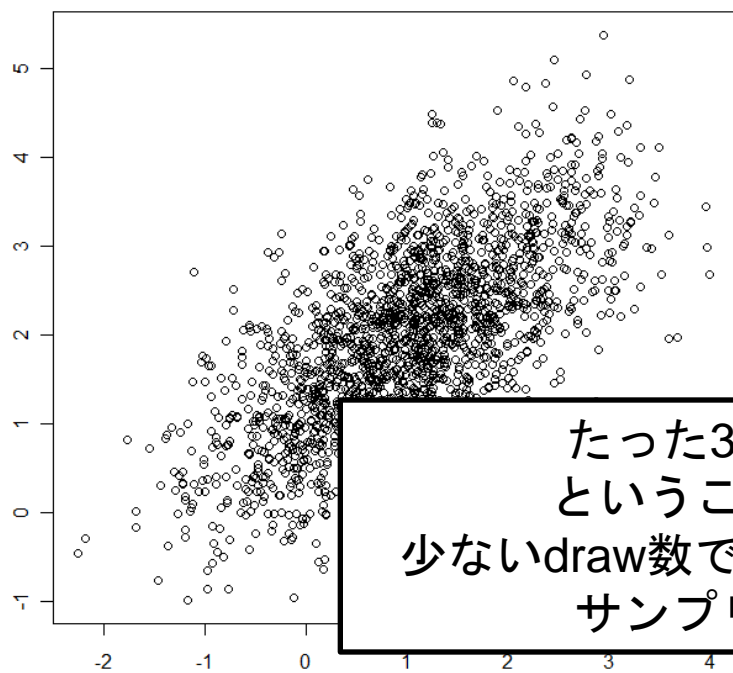
$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right)$$

Rの多変量正規分布の乱数発生関数から2000個

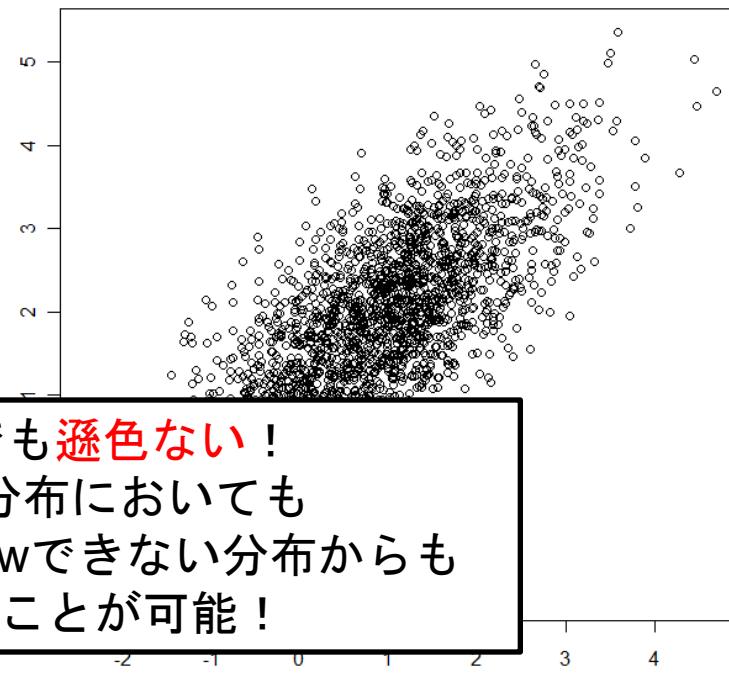
$$x_1 \sim N(1 + 0.7(x_2 - 2), 1 - 0.7^2)$$

$$x_2 \sim N(2 + 0.7(x_1 - 1), 1 - 0.7^2)$$

ギブスサンプラー(3000draw 1000はburn-in)



x.positive2[,1]



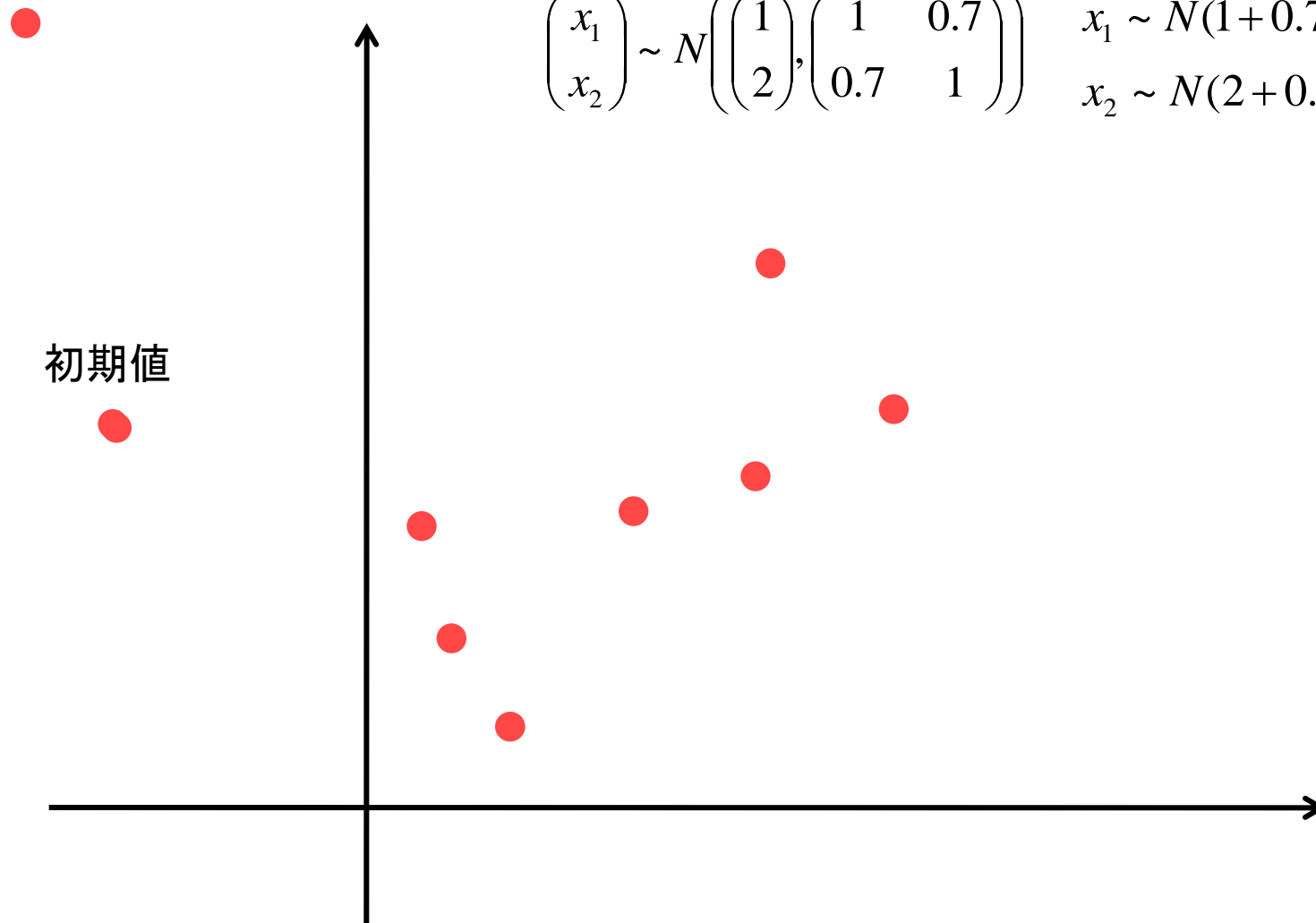
x.gibbs[,1]

たった3000drawでも遜色ない！
 ということは他の分布においても
 少ないdraw数で簡単にdrawできない分布からも
 サンプルングすることが可能！

ギブス・サンプラーのイメージ

$$\begin{pmatrix} x_1 \\ x_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 1 \\ 2 \end{pmatrix}, \begin{pmatrix} 1 & 0.7 \\ 0.7 & 1 \end{pmatrix}\right) \quad \begin{aligned} x_1 &\sim N(1+0.7(x_2-2), 1-0.7^2) \\ x_2 &\sim N(2+0.7(x_1-1), 1-0.7^2) \end{aligned}$$

初期値



メトロポリス・ヘイスティングス法

- メトロポリス・ヘイスティングス法は「マルコフ連鎖的棄却サンプリング」
- 棄却サンプリングと同様，提案分布 $q(x)$ からサンプリングすることを考える．ここで重点サンプリングと同様に，重み付き分布の考え方を使って，事後分布を重要度重みで置き換える．
- そして重要度重み（インポートランス比）の比（ここがマルコフ的）と一様乱数を比べて採択するかどうかを選ぶ
- 採択されなかった場合，単純に棄却してやり直すのではなく $t-1$ 期をそのまま用いる．
- これぞMH法だ！

メトロポリス・ヘイスティングス法

1. 初期値 $z^{(0)}$ を決め, $t=1$ とおく
2. 現在, $z^{(t-1)}$ であるとき, 次の値 $z^{(t)}$ の候補 z' を提案分布 $q(z^{(t-1)}, z)$ から発生させ

$$\alpha(z^{(t-1)}, z^*) = \min \left\{ \frac{\tilde{p}(z^*)}{\tilde{p}(z^{(t-1)})} \frac{q(z^* | z^{(t-1)})}{q(z^{(t-1)} | z^*)}, 1 \right\}$$

と定義する

3. $(0,1)$ 上の一様乱数 u を発生させて,

$$z^{(t)} = \begin{cases} z^* & \text{if } U \leq \alpha(z^{(t-1)}, z^*) \\ z^{(t-1)} & \text{if } U > \alpha(z^{(t-1)}, z^*) \end{cases}$$

4. t を $t+1$ として, 2. に戻る

スライス・サンプリング

■ むり

ここまでのまとめ（３）

- ギブスサンプリング
 - 条件付き確率が書き下せればかなり効率良い
 - 1つ1つ固定していったうえでのサンプリング
- メトロポリスヘイスティングス法
 - 重点サンプリングの重み付け比を用いて事後分布に近似
 - 計算に時間がかかる（提案分布の選び方がやはり重要）
- スライスサンプリング
 - むり

まとめ

- not MCMCなサンプリングからみていくことでMCMCなサンプリングの理解も深まったはず
- MCMCは万能ではなく，計算時間がかかったり，相関をもってしまうたりする
- 使い道が重要