本レジュメの巻末には、Appendix のページを付けている。必要に応じて参照する。

11 章前書き

まずは本章で学ぶサンプリング法の位置づけについて確認したい。この章で取り扱う問題は、ある関数 $f(\boldsymbol{z})$ の確率分布 $p(\boldsymbol{z})$ のもとでの期待値の計算に関することである。これを定式化すると、もし $\boldsymbol{z}$ が連続変数の場合は

$$E[f] = \int f(\boldsymbol{z})p(\boldsymbol{z})d\boldsymbol{z} \tag{1}$$

を評価するということになる。そして今回、この式 1 において、$E[f]$ が解析的に求められないとする（See 図 11.1)。

そこでサンプリング法の出番となる。サンプリング法の考え方は次の通りである。

・分布 $p(\boldsymbol{z})$ から独立に抽出されたサンプルの集合 $\boldsymbol{z}^{(l)}$ $(l = 1, \cdots, L)$ を得る

・この $\boldsymbol{z}^{(l)}$ を用いて、式 1 を以下の通り有限和で近似する

$$\hat{f} = \frac{1}{L}\sum_{l=1}^{L} f(\boldsymbol{z}^l) \tag{2}$$

そしてこの期待値 $E[\hat{f}]$ は、サンプルが分布 $p(\boldsymbol{z})$ に従って抽出される限り、$E[\hat{f}] = E[f]$ が成立する。したがい、推定量 $\hat{f}$ は正しい（真の）平均を持つ。また、$\hat{f}$ の分散については、式（11.3）のように表すことができる（See App.1)。比較的少数のサンプルで高い精度を確保できる、のところがわからなかった。$\boldsymbol{z}$ の次元数の話を持ち出していたけど、それと精度の関係性がよくわからなかった。

グラフィカルモデルとの関係性からサンプリング戦略にかけての話は理解不能。読み合わせで解読させてほしい。

1　基本的なサンプリングアルゴリズム

与えられた分布からランダムにサンプルを生成したい。そのためには乱数の生成が必要となり、ここでは、その乱数生成のアルゴリズムについて学んでいく[*1]。

1.1　標準的な分布

ここでは、サンプリング法の代表的な手法である「変換法」について学ぶ[*2]。

単純な非一様分布から乱数を生成するということがやりたい。そのためには、一様分布の乱数の発生源が必要であり、これは手元にあると仮定しよう。いま z が区間 $(0, 1)$ で一様に分布し、ある関数 $f(\cdot)$ を用いて z の値を $y = f(z)$ のように変換することを考える[*3]。このとき y の分布は、ヤコビアンを導入することで

$$p(y) = p(z)|\frac{dz}{dy}| \tag{3}$$

と表される。

[*1] 「自然科学の統計学」第 11 章において、乱数生成を網羅的に取り扱った。
[*2] 「自然科学の統計学」で、我々が「逆関数法」として学んだものである。
[*3] 変換後の y についての分布が、冒頭で「単純な非一様分布」と設定したものである。

ここでポイントになるのは、変数変換のための $f(\cdot)$ をどうやって選んでくるか、ということである。z に変換を施し得られた y が、求めたい特定の分布 $p(y)$ に従っている必要がある。そのためには以下の手続きを辿ればよい。

・求めたい分布 $p(z)$ の累積分布関数（以下 c.d.f.）を計算する
・一様分布に従う z に対して、上で求めた c.d.f. の逆関数を食わせる[*4]

それでは例えば、求めたい分布が次のような指数分布であるときに、変換法を適用してみよう。

$$p(y) = \lambda \exp(\lambda y) \tag{4}$$

いま $p(y)$ の c.d.f. を計算し、その逆関数を求めることにする（See App.2)。

同様にして、求めたい分布が以下の関数である場合に、変換に用いる関数の具体形を導出する。

・コーシー分布（See App.3）
・2 変量ガウス分布（See App.4）

最後に変換法の注意点を述べ、次サブセクションへの橋渡しをおこなう。変換法を可能にする要素は以下の 2 点である。

・求めたい分布の c.d.f. が解析的に求められること
・計算した c.d.f. の逆関数が解析的に求められること

実は、これら 2 点が常に満たされることはない。そこで、より一般的に使える手法として、棄却サンプリングという手法を次のサブセクションで学ぶことにする。

1.2 棄却サンプリング

ここでは、変換法よりも一般的な状況下でも適用できる「棄却サンプリング」について学ぶ。簡単のため一変数の場合で議論をおこなう。

いま単純で標準的ではない分布[*5] $p(z)$ からのサンプリングをおこないたい。$p(z)$ に対する細かい条件は、教科書を参照されたい。ポイントとしては、その $p(z)$ からのサンプリングが今はできないと考えよう。

そこで、提案分布と呼ばれる、簡単にサンプリングができるような分布 $q(z)$ および、その定数倍で表現される比較関数 $kq(z)$ を導入する。これらを導入し、元の分布 $\tilde{p}(z)$（ただし規格化されていない）との大小関係に制約を設けることで、棄却サンプリングを実行することができるようになる。その実行手続きは次の通りである。ただし、図 11.4 を参照せよ。

・サンプル z_0 を決める、同時に $kq(z_0)$ が決まる
・u_0 を区間 $[0, kq(z_0)]$ における一様分布から決める
・z_0 を受理するか棄却するか次のルールにもとづいて決める
・ $u_0 < \tilde{p}(z)$ なら受理
・ $u_0 > \tilde{p}(z)$ なら棄却

ところで、サンプルが生成される確率が $q(z)$ であり、そのサンプルが受理される確率が $\tilde{p}(z)/kq(z)$ である

[*4] ビジュアル的な理解をするためには「自然科学の統計学」321 ページを読むことをオススメする。
[*5] 「単純で標準的ではない」の意味は、求めたい分布の c.d.f. および、その c.d.f. の逆関数が解析的に書けないということと理解した。

ことから、結局サンプルが生成〜受理されるまでの確率は

$$\begin{aligned} p(\text{受理}) &= \int \{\tilde{p}(z)/kq(z)\}q(z)dz \\ &= \frac{1}{k}\int \tilde{p}(z)dz \end{aligned} \tag{5}$$

となる。

なお、サンプルが受理される確率を大きくするためには、定数 k を小さくし、棄却域を狭くしてやればよい[*6]。素朴な疑問だが、受理されることは良いことなのだろうか？

棄却サンプリングの利用として、求めたい分布がガンマ分布であるケースについて考えよう。明らかにガンマ分布は c.d.f. を解析的に求めることができない。そこで、提案分布としてコーシー分布を採用することにしよう。コーシー分布はガンマ分布に形状が近いこと、さらに変換法が適用できる分布であることを満たしている。

ここで、提案分布と元の分布（規格化されていない）との大小関係についての制約を満たすために、コーシー分布を一般化する必要がある。そのためには一様分布の確率変数 y を $z = b\tan y + c$ を用いて変換してやればよい。これによって乱数は

$$q(z) = \frac{k}{1 + (z-c)^2/b^2} \tag{6}$$

に従って分布することが導ける（See App.5）。ここで比較関数と元の分布（規格化されていない）との位置関係を図 11.5 にて確かめる。

最後に補足すると、実は提案分布 $q(z)$ のような包絡分布は常に解析的に書けるわけではない。そこで次回は、そのような場合でもサンプリングを可能にする手法として適応的棄却サンプリングを学ぶ。

[*6] グラフを描けばすぐわかる。$kq(z)$ が $\tilde{p}(z)$ より小さい領域が存在してはいけない、という制約を考えて、グラフ平面において $kq(z)$ を z 軸に近づけたときの振る舞いを見ればよい。