

# 一般化相対二乗誤差に基づく低確率事象強調サンプル法

A Sampling Method based on Generalized Relative Square Error  
to Emphasize Low Probability Events

中村文美 <sup>\*1</sup>

Tomomi Nakamura

長谷川博 <sup>\*1</sup>

Hiroshi Hasegawa

鷲尾隆 <sup>\*2</sup>

Takashi Washio

<sup>\*1</sup>茨城大学大学院理工学研究科

Graduate School of Science and Engineering, Ibaraki University

<sup>\*2</sup>大阪大学産業科学研究所

The Institute of Scientific and Industrial Research, University of Osaka

A method of data sampling from a huge data set is discussed. We introduce a generalized relative square error to emphasize low probability events and figure out the best sampling weight to reduce the error. Our arguments are based on the large deviation theory. Large reduction in the generalized relative square error was numerically confirmed for the best sampling weight. We also propose to use Wang-Landau algorithm in data sampling. This algorithm is not only efficient to estimate a distribution of the original data, but also useful in data sampling to suppress the statistical errors.

## 1. 背景

近年、ネット社会の発達に伴い、カードや携帯電話が普及し、時々刻々の個人レベルの情報が収集されている。また気象などの解析のための大規模シミュレーションなどによっても、膨大なデータが生み出されている。この状況から、ペタ（エクサ）というスケールの大規模データから有益な情報を導かなければならぬ状況になりつつある。

しばしば、このような大規模データの全てを用いて解析することは、効率的ではないので、サンプル・データを用いて解析することが考えられる。しかしながら、ただ一様にサンプリングすれば、頻出事象のデータは取り込めて、低確率事象のデータは捨てられてしまう可能性が高い。頻出事象と低確率事象が同程度の重要性を持つ場合は、一様サンプリングで良いが、大規模事故や大規模災害のように、滅多に起きない事象の方が重要な場合がある。このような場合、ハインリッヒの法則を考慮すれば、頻出事象から低確率事象まで、切れ目なくサンプリングする必要がある。この論文では、低確率事象を考慮したサンプリング法を提案する。

最初、低確率事象の重要性を考慮して一般化した相対二乗誤差を導入する。次に大偏差理論 [Ellis 06] を基礎に、一般化相対二乗誤差最小を実現するサンプル分布を特定する。そして、特定したサンプル分布に対して、一般化相対二乗誤差が実際小さくなることを数値実験により確認した結果を報告する。最後に、Wang-Landau 法 [Wang-Landau 01] を適用して、データ分布が不明のデータからデータ分布を推定しつつ、サンプリングを行う場合の一般化相対二乗誤差の減少について議論する。

## 2. 一般化相対二乗誤差最小分布

実数  $x$  は定義域  $[x_{\min}, x_{\max}]$  に含まれるとし、大規模データが従う確率分布を  $f(x)$  とする。確率分布  $f(x)$  から生成された大数  $N_L$  のデータ集合を  $X = \{x_n | n = 1, \dots, N_L\}$  とする。定

連絡先: \*1 e-mail: 16nm110t@vc.ibaraki.ac.jp,

hiroshi.hasegawa.sci@vc.ibaraki.ac.jp

2 e-mail: washio@ar.sanken.osaka-u.ac.jp

義域を  $i_{\max}$  個の区間  $I_i = [x_{i-1}, x_i]$  ( $i = 1, \dots, i_{\max}$ ,  $x_{\min} = x_0, x_{\max} = x_{i_{\max}}$ ) に分割し、 $i$  番目の区間  $I_i$  に含まれる確率を  $f_i$  とする<sup>\*1</sup>。データ集合  $X$  からデータ  $x \in I_i$  をランダムにサンプルした時、重み  $\pi_i$  で選択するとする。この重み  $\pi_i$  を低確率事象に配慮して設計する。

区間  $I_i$  にデータがサンプルされる確率は  $p_i = \pi_i f_i$  で与えられる。しかし統計ゆらぎのため、実際に区間  $I_i$  にデータがサンプルされる数  $n_i$  は、一般には  $n p_i$  からずれる。ここで  $n$  は全サンプル数とする。今、確率  $q_i = n_i/n$  と定義するとき、確率分布  $p, q$  を、各々  $p = \{p_i | i = 1, \dots, i_{\max}\}$ ,  $q = \{q_i | i = 1, \dots, i_{\max}\}$  とする。このとき、期待されるサンプル分布  $p$  に対して、サンプルでの実現分布  $q$  の出現確率は、大偏差理論 [Ellis 06] より、以下のように与えられる。

$$P(q|p) \approx \exp[-n D_{KL}(q : p)] \quad (1)$$

ここで、カルバック・ライブラー (KL) ダイバージェンス  $D_{KL}(q : p) = \sum_i q_i \log(q_i/p_i)$  は  $q$  と  $p$  の情報的距離を与える。 $q$  と  $p$  が等しい時の  $D_{KL}(q : p) = 0$  となる。

期待されるサンプル分布  $p$  が与えられたときの、誤差の期待値を以下のように定義する。ここで低確率事象を考慮して一般化した相対二乗誤差を用いる。

$$E_{\nu}(p) = \int \sum_{i=1}^{i_{\max}} \left| \frac{f_i - \hat{f}_i(q|p)}{f_i^{\nu}} \right|^2 P(q|p) \delta(\sum_{j=1}^{i_{\max}} q_j - 1) dq_1 dq_2 \dots dq_{i_{\max}} \quad (2)$$

ここで  $\hat{f}_i(q|p) = f_i q_i / p_i = q_i / \pi_i$ 、 $\delta$  関数は確率の規格化のために導入してある。 $\nu$  は低確率事象を考慮して選択する実数とする。 $\nu = 0$  のとき、 $E_{\nu}(p)$  は通常の二乗誤差、 $\nu = 1$  のとき、 $E_{\nu}(p)$  は通常の相対二乗誤差を与える。例えば事故データにおいては、大事故の確率は小さくとも、甚大な被害を与える。被害額を考慮して確率の小さい大事故をより優先させるサンプリングでは、 $\nu > 1$  が選択される。

一般化相対二乗誤差  $E_{\nu}(p)$  を最小にするサンプル分布  $p$  を導出する。まず  $P(q|p) \approx \exp[-n D_{KL}(q : p)]$  を見積もる。 $n$

\*1 区間選択としては、簡単な均等分割でも良いが、区間幅に重みをつけた区間選択も可能である。

が十分に大きいとき、 $D_{KL}(q : p) \approx 0$ 、すなわち  $q \approx p$  の近傍のみが積分に寄与できる。そこで  $q_i = p_i + \Delta q_i$  として、 $\Delta q_i$  の 2 次まで見積もると、 $D_{KL}(q : p) \approx \sum_i \Delta q_i^2 / (2p_i)$  となり、一般化相対二乗誤差  $E_\nu(p)$  は以下のように書き換えられる。

$$E_\nu(p) \approx C \int \sum_{i=1}^{i_{\max}} \frac{f_i^{2(1-\nu)} \Delta q_i^2}{p_i^2} \exp[-n \sum_{j=1}^{i_{\max}} \frac{\Delta q_j^2}{2p_j}] \\ \times \delta(\sum_{j=1}^{i_{\max}} \Delta q_j) d\Delta q_1 d\Delta q_2 \dots d\Delta q_{i_{\max}} \quad (3)$$

ここで  $C$  は規格化定数である。更に、 $\delta$  関数の積分表示、 $\delta(x) = \int_{-\infty}^{\infty} \exp[-ikx] dk / (2\pi)$  を上の一般化相対二乗誤差  $E_\nu(p)$  の式に代入して、鞍点法で  $\Delta q_i^2$  までガウス積分で見積もると、

$$E_\nu(p) \approx \frac{1}{n} \sum_{i=1}^{i_{\max}} f_i^{2(1-\nu)} \left( \frac{1}{p_i} - 1 \right) \quad (4)$$

となる。ガウス積分であるが、実際の計算はかなり面倒である。

次に、一般化相対二乗誤差  $E_\nu(p)$  を最小にするサンプル分布  $p^{(\nu)}$  を導出する。 $p^{(\nu)}$  は規格化条件  $\sum_i p_i = 1$  の下での一般化相対二乗誤差  $E_\nu(p)$  最小化の式の解として求めることができ、

$$p_i^{(\nu)} = \frac{f_i^{1-\nu}}{\sum_j f_j^{1-\nu}} \quad (5)$$

となる。 $E_\nu(p)$  が通常の二乗誤差 ( $\nu = 0$ ) のときは、 $p_i^{(0)} = f_i$  すなわち  $\pi_i = 1$  となって、単純サンプリングが二乗誤差最小のサンプリング法になる。 $E_\nu(p)$  が通常の相対二乗誤差 ( $\nu = 1$ ) のときは、サンプル分布が一様分布  $p_i^{(1)} = 1/i_{\max}$  ( $\pi_i = 1/(i_{\max} f_i)$ ) となって、全ての事象を一様にサンプルすることが相対二乗誤差を最小にする。

最後に、この後の数値実験の準備として、サンプル分布  $p^{(\eta)}$  に対する一般化相対二乗誤差  $E_\nu(p^{(\eta)})$  を導出しておく。

$$E_\nu(p^{(\eta)}) \approx \frac{1}{n} \left\{ \sum_i f_i^{1-2\nu+\eta} \sum_j f_j^{1-\eta} - \sum_i f_i^{2(1-\nu)} \right\} \quad (6)$$

### 3. 数値実験

人工データからサンプリングを行って、サンプル分布  $p^{(\eta)}$  ( $\eta = 0, 1/2, 1, 2$ ) を生成し、一般化相対二乗誤差  $E_\nu(p^{(\eta)})$  ( $\nu, \eta = 0, 1/2, 1, 2$ ) を計算し、理論の結果と比較した。またサンプル分布  $p^{(\nu)}$  に対する  $E_\nu(p^{(\nu)})$  が、他のサンプル分布に対する  $E_\nu(p^{(\eta)})$  ( $\eta \neq \nu$ ) に比較して小さくなることを、数値実験により確認した。具体的には、以下の数値実験を行った。

1. Box-Muller 法で 200 万個の標準正規乱数を生成して、元データとした。
2.  $[-3, 3]$  を定義域として、均等幅の 100 個の区間でヒストグラムを作成し、データ分布  $f_i$  ( $i = 1, 2, \dots, 100$ ) を導出した。
3. 導出したデータ分布  $f_i$  から、理論値  $E_\nu(p^{(\eta)})$  ( $\nu, \eta = 0, 1/2, 1, 2$ ) を計算した。
4.  $\eta = 0, 1/2, 1, 2$  のサンプル分布  $p^{(\eta)}$  は、元データから重み  $\pi_i \propto f_i^{-\eta}$  で選択して生成した。サンプル分布のデータ数は 1000 個として、一般化相対二乗誤差  $E_\nu(p^{(\eta)})$  ( $\nu = 0, 1/2, 1, 2$ ) を計算した。統計サンプル数は 1 万とした。

数値実験の結果を表 1 に示す。理論と数値実験が良い精度で完全に一致した。特に一般化相対二乗誤差  $E_\nu(p^{(\eta)})$  に対して

は、期待通り  $\nu = \eta$  の場合に誤差が最小となった。例えば、相対二乗誤差について、重み無しでサンプルした場合と最小を実現する一様分布になるようにサンプルした場合と比較すると、相対二乗誤差は約 1/5 に減少した。

表 1. サンプル分布  $p^{(\eta)}$  に対する一般化相対二乗誤差  $E_\nu(p^{(\eta)})$

$E_\nu(p^{(\eta)})$	$\eta = 0$	$\eta = 1/2$	$\eta = 1$	$\eta = 2$
$\nu = 0$ 実験	9.86E-4	1.10E-3	1.68E-3	1.65E-3
$\nu = 0$ 理論	9.86E-4	1.10E-3	1.68E-3	1.65E-3
$\nu = 1/2$ 実験	0.0993	0.0772	0.0987	0.844
$\nu = 1/2$ 理論	0.0990	0.0772	0.0990	0.843
$\nu = 1$ 実験	50.5	15.5	9.90	49.6
$\nu = 1$ 理論	49.6	15.4	9.90	49.6
$\nu = 2$ 実験	2.25E8	3.80E7	8.65E6	2.38E6
$\nu = 2$ 理論	2.16E8	3.72E7	8.56E6	2.56E6

実用においては、元のデータ分布が不明の場合が想定される。その場合に備えて、Wang-Landau 法 [Wang-Landau 01] を適用して、データ分布を効率的に推定すると同時にサンプリングを行うことを試みた。今回、Wang-Landau 法は、元データのデータ分布の効率的推定に有用であるだけでなく、サンプリングの統計誤差の抑制にも効果があることが明らかになった。簡単のために、目標サンプル分布が一様の  $p^{(1)}$  場合に、Wang-Landau 法を用いてサンプリングを行ったところ、相対二乗誤差  $E_1(p^{(1)})$  は更に 1/5 以下に減少した。

### 4. 結論

この論文で、我々は、大規模実データからの低確率事象強調サンプリングに、一般化相対二乗誤差最小を実現するサンプル分布を用いることと同時に、元データのデータ分布の推定とサンプリングに、Wang-Landau 法を用いることを提案した。我々は、低確率事象を強調して扱うために、相対二乗誤差をより一般化した一般化相対二乗誤差を導入し、誤差最小を実現するサンプル分布を推定した。この結果は、数値実験により、良い精度で確認された。更に Wang-Landau 法は、実用において必要な元データ分布の効率的推定に有用であるだけでなく、サンプリングにおいても、統計ゆらぎの抑制効果があることを数値実験で確認した。

この論文の基礎となる着想は、マルチカノニカルマルコフ鎖鎖モンテカルロ法から得ている [Berg-Neuhaus 92]。そこでは低確率事象を強調するために、サンプルヒストグラムのフラット化、すなわちすべての事象を一様にサンプルするためには、状態密度の逆数の重みでサンプリングを行う。これは我々の相対二乗誤差最小の一様サンプル分布に対応する。また状態密度の効率的な計算法として開発されたのが、Wang-Landau 法である。

### 参考文献

- [Ellis 06] Ellis, R. S.: Entropy, Large Deviations and Statistical Mechanics, Springer Publication (2006).
- [Wang-Landau 01] Wang, F. and Landau, D. P.: Phys. Rev. Lett. **86** 2050–2053 (2001).
- [Berg-Neuhaus 92] Berg, B. and Neuhaus, T.: Phys. Rev. Lett. **68** 9–12 (1992).