論文の内容の要旨

Efficient Analysis of Event Data and Application for Optimal Termination: Practice in Inelastic Neutron-Scattering Experiments
（イベントデータの効率的な分析と最適停止への応用：非弾性中性子散乱実験での実践）

氏　名　　武藤　健介

1. はじめに

近年，数値シミュレーションや計測技術の向上により，様々な分野で多次元かつ大容量のデータが得られている．これらのデータには，ノイズの低減や，データ容量の削減を目的とした前処理がしばしば行われる．現状，ほとんどの前処理は専門家が手動で行なっており，人的コストや人によって処理の基準が一定でない事が問題である．そこで，これらの前処理を自動化するアルゴリズムの開発が求められている．

中性子散乱実験，放射線計測，神経科学でのスパイク計測などイベントデータを取得する実験が多数存在する．本研究では，特に非弾性中性子散乱実験に着目する．中性子は電荷を持たないので，物質中の電子との相互作用が無く，透過率が高いことが知られている．そのため，物質の内部構造や元素分布を決定できる．また，水素やリチウムなど，軽元素の位置や構造の特定にも適している．非弾性中性子散乱実験では，個体中の原子間のダイナミクスを調べることができるので，新物質の開発などに期待されている．近年，大強度陽子加速器施設物質・生命科学実験施設(J-PARC/MLF) のチョッパー分光器 4SEASONS により[1]，高効率に大容量の非弾性中性子散乱実験データを得られるようになった．データは，散乱前後でのエネルギー（1 次元）と運動量（3 次元）の変化を表す，4 次元空間中の点群として得られる．計測が独立である場合，イベントの観測過程はポアソン分布に従うとモデル化でき，データは確率的な揺らぎを持って観測される．観測データに含まれるノイズの低減とデータ容量の削減のため，一般にイベントデータに対してヒストグラムやカーネルによる平滑化が行われている．本研究では，現在 J-PARC で主に扱われているヒストグラムに注目する．ヒストグラムはイベントデータに限らず，データを平滑化してノイズの影響を低減しながら可視化する手法の一つである．ヒストグラムを作成する際には，データを矩形の集合として表す．この矩形は横幅（ビン幅）をチューニングパラメータとし，縦の長さを横幅の範囲に含まれる数量の合計として表現する．ここで，ビン幅を大きくすると平滑化によってイベントデータの確率的なゆらぎを低減する事ができるが，一方でデータがもともと持っていた構造は見えなくなるというトレードオフの問題があり，この設定は自明ではない． J-PARC では Utsusemi と呼ばれるソフトウェアを用いて[2]，4 次元データからスライス幅を設定して一部の 2 次元データを可視化する．そして，残り 2 次元のビン幅の調節を行なっている．現状の前処理にはビン幅を各研究者が恣意的に設定するという問題と，高次元データから低次元のスライスを切り出す際に特定の軸方向のスライス幅を恣意的に設定する問題がある．切り出すスライス幅とその場所は，切り出したデータに対するビン幅の決定に影響することも予想される．また 3 次元以上のデータを目視で処理するのは非常に高コストである．そこで本研究では，多次元のイベントデータに対するヒストグラムのビン幅最適化手法を提案する．

提案手法の効果的な応用先として，データの分析だけでなく，計測の打ち切り判定が考えられる．多くの実験施設では，得られたデータが十分であるか判断する基準がなく，計測が冗長化す

る傾向がある．大量のデータを得ることでデータの持つ微細な特徴を抽出できるが，装置の分解能を超えた情報の抽出は不要である．そこで我々は，計測を行いながら得られたデータに対してヒストグラムの最適ビン幅を計算し，それが装置の分解能を下回る時に計測を停止する手法を提案する．ここで，データ数が増える毎に最適ビン幅が小さくなることが知られている[3, 4]．最適ビン幅の計算は並列化可能なため，リアルタイムでの打ち切り判定に適している．さらに，我々はベイズ最適化（BO）を用いた効率的な最適ビン幅の探索手法を提案する．BO はしばしば最適化問題を解くために用いられる探索手法である[5]．BO は，ガウス過程（GP）により得られた予測分布をもとに，最適値探索の効率化を図る．ここで，GP は信頼度付きでデータ点の内挿を行う手法として知られている[6]．最適ビン幅探索の更なる効率化のために，数時間前の停止判定での計算結果を用いる工夫が考えられる．そこで，我々は過去の停止判定の際に計算した情報を組み込んだベイズ最適化を提案する．

論文の構成に関して，図 1 に示すように，2 章で多次元ヒストグラムのビン幅最適化について述べる．そして，3 章と 4 章でイベント計測の停止判定手法の提案とその効率化について述べ，5 章で結論を述べる．

2. 多次元ヒストグラムのビン幅最適化

2 章では多次元ヒストグラムのビン幅最適化に関して述べる．ビン幅最適化の先行研究としては，単位時間あたりのニューロンの発火頻度を表す発火率をヒストグラムの形で推定する為に，1 次元のイベントデータからヒストグラムのビン幅最適化を行う手法が知られている[3]．多数の試行により得られたニューロンのスパイク時系列に関して，特定の時間幅に含まれるスパイクカウントはポアソン分布に従うことが知られている．また同様に，特定の領域内での，中性子カウントはポアソン分布に従うことが知られているため，非弾性中性子散乱実験のデータに適用可能な手法開発を試みた．具体的には，ビン幅最適化手法の多次元拡張を試みた．

1 次元のビン幅最適化手法[3]に関して，ビン幅最適化のコスト関数について述べる．ビン幅最適化は密度推定の問題として定式化される．イベントデータの従う真の確率分布をヒストグラムの形で推定する問題である．即ち，ヒストグラムと真の分布のズレを記述し，その最小化問題として定式化できる．ここでは，真の分布からのズレを表す量として，平均積分自乗誤差（MISE）を用いる．推定性能の評価基準として，カルバック・ライブラー（KL）ダイバージェンスも知られているが，ヒストグラムにおける確率密度の推定においては，KL ダイバージェンスが発散しうる．従って，本稿では MISE を用いる．MISE からビン幅に依存する項のみを抽出することで，コスト関数を得る．最適ビン幅の探索に関して，一般的にはグリッドサーチによって，コスト関数を最小化するビン幅を見つける．アルゴリズムを多次元化する上で，次元数に対して指数関数的に計算量が増大する．コスト関数の多次元拡張は容易に行えるが，計算量の削減が実運用に向けた喫緊の課題である．そこで，我々はコンピュータビジョンなどで用いられる，Summed-area tables（SAT）と呼ばれる手法に目をつけた[7]．SAT は累積和を計算する前処理を行うことで，計算コストを大幅に削減する．これにより，4 次元のビン幅最適化が現実的な計算量の問題として扱えるようになった．

本章では，提案手法の挙動を調べるため，人工データを用いた数値実験を行なった．まずは，現状実験者が 4 次元データから，位置とスライス幅を指定して切り出した 2 次元データに対して，目視で処理を行なっている点に注目する．我々は，4 次元の人工データに対して，スライスして切り出した 2 次元データと，4 次元データ全体に対するビン幅最適化の結果を比べた．結果として，得られた最適ビン幅は両者で異なりうることが分かった．スライスして切り出す際に，平面上に含まれる中性子カウントが変化し，それがビン幅の決定に大きく影響を与える．そのため，4 次元データ全体に対して，提案手法を適用するのがよいと言える．次に，実データを想定し，データ数やバックグラウンドノイズの大きさを変化させた状況での提案手法の挙動を調べた．結果

として，ノイズが大きい場合やデータ数が少ない場合は，ゆらぎを低減する効果が優先され，広いビン幅が選択されることが分かった．

3. イベント計測の最適停止

3 章では，まず 4SEASONS で得られた実データに対してビン幅最適化を行なった．結果として，最適ビン幅が装置の分解能を大きく下回った．次に，実験を行う過程でイベント数が増加する状況を想定し，データをダウンサンプリングして，イベント数を変化させてビン幅最適化を行なった．その結果，イベント数を半分にした状況で，既に最適ビン幅が装置の分解能を下回っていることが分かった．このように，多くの実験施設には計測の停止基準がないため，計測が冗長化する傾向がある．大量のデータを取得することで，観測対象に関する微細な情報を抽出することができる．しかし，実際の計測には装置の分解能が存在するので，それを超えた観測は不要である．そこで，多次元ビン幅最適化を応用した停止基準を提案する．我々は，イベントデータに対して計算した最適ビン幅が，データの持つ解像度に対応すると考えた．つまり，実験者は計測を行いながら得られたデータに対してヒストグラムの最適ビン幅を計算し，それが装置の分解能を下回る時に計測を停止すればよい．ここで，イベント数の増加に伴い，最適ビン幅が単調に減少することが知られている．また，停止判定を行う上で，イベントの従う確率分布の定常性が保証されている必要がある．4SEASONS では，近年の実験技術の向上によって，イベント計測の定常性が担保されている[8].

ビン幅最適化はコスト関数の最小化問題として位置付けられる．現状グリッドサーチを行うことで，最小値を探索しているが，リアルタイムでの停止判定に向けた探索の効率化を実現したい．ここで，コスト関数は並列計算可能であるが，膨大な計算資源を必要とするので，アルゴリズムによる探索の効率化を目指す．我々は，コスト関数の局面が滑らかではあるが，いくつかの局所解を持つ点に着目した．ここから，勾配法ではなく，内挿を行いながら次の探索点を選択する BO が有効だと考えた．BO は，GP を用いて得られた予測分布をもとに，次の探索点を決定する．ここで，GP は信頼度付きでデータ点の内挿を行う手法として知られている．BO で次の探索点を選択する際に，獲得関数と呼ばれる指標を用いる．これは，探索と情報の活用のトレードオフを表すような関数で，予測分布を用いて計算される．数値実験の結果，ベイズ最適化の有効性が示された．特に，イベント数が多いほど顕著に探索効率の向上が見られた．

4. 停止判定の効率化

4 章では，最適ビン幅探索の更なる効率化を目指す．具体的には，数時間前の停止判定でのコスト関数の計算結果を用いる方策を検討する．島崎と篠本によって，コスト関数のデータ数に対する外挿の定式化が行われている．これを用いることで，過去に計算したコスト関数を用いて，データ数が増えた状況でのコスト関数の値を推定することができる．我々は，ベイズ最適化に用いる予測分布に，外挿されたコスト関数の情報を組み込む方法を提案する．具体的には，外挿されたコスト関数から計算した GP の予測分布と，データ数が増えた状況でのコスト関数の計算結果に GP を適用して得た予測分布の同時分布を扱う．この同時分布に対して BO の獲得関数を計算すればよい．ここで，外挿されたコスト関数から計算した GP の予測分布をコスト関数の事前分布と呼ぶことにする．我々は事前分布の寄与度合いを調節するハイパーパラメータ（HP）を 1 つ導入した．数値実験の結果，大幅な探索効率の向上が見られた．また，提案手法は HP に対してロバストだということが分かった．

参考文献

[1] R. Kajimoto, M. Nakamura, Y. Inamura, F. Mizuno, K. Nakajima, S. Ohira-Kawamura, T. Yokoo, T. Nakatani, R. Maruyama, K. Soyama, et al., J. Phys. Soc. Jpn80, SB025 (2011).
[2] Y. Inamura, T. Nakatani, J. Suzuki, and T. Otomo, J. Phys. Soc. Jpn. 82, SA031 (2013).

[3] H. Shimazaki and S. Shinomoto, Neural Computation19, 1503 (2007).
[4] K. Muto, H. Sakamoto, K. Matsuura, T. Arima, and M. Okada, J. Phys. Soc. Jpn88, 044002 (2019).
[5] J. Mockus, V. Tiesis, and A. Zilinskas, Towards global optimization2, 2 (1978).
[6] C. K. Williams and C. E. Rasmussen, Vol. 2 (MITpress Cambridge, MA, 2006).
[7] F. C. Crow, inProceedings of the 11th annualconference on Computer graphics and interactive techniques (1984) pp. 207–212.
[8] Kajimoto, R., Nakamura, M., Inamura, Y., Kamazawa, K., Ikeuchi, K., Iida, et al., J. Phys.: Conf. Ser (Vol. 1021, p. 012030) (2018).

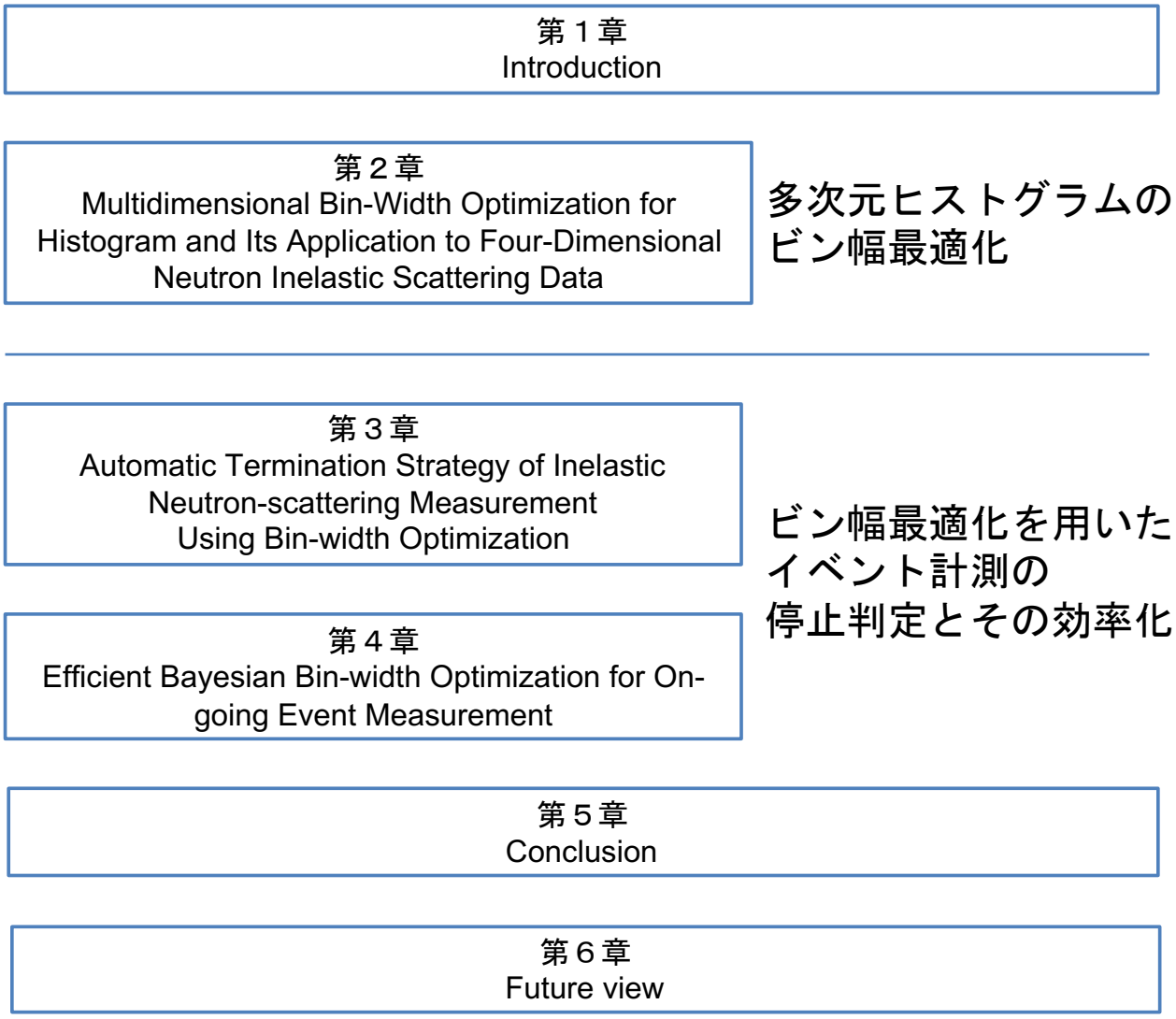

図１：本論文の構成.