Integrated Gradients における理想の積分ステップ数はインスタンス毎に異なる

牧野 雅紘 [1]　浅妻 佑弥 [1,2]　佐々木 翔大 [3,1]　鈴木 潤 [1,2]
[1] 東北大学　[2] 理化学研究所　[3] 株式会社サイバーエージェント
{masahiro.makino.r6, asazuma.yuya.r7}@dc.tohoku.ac.jp
sasaki_shota@cyberagent.co.jp　jun.suzuki@tohoku.ac.jp

概要

Integrated Gradints (IG) は、機械学習モデルの挙動を説明する手法である。IG は公理により値の加法性が保証される利点を持つが、実際は数値積分の誤差から公理が保証されない可能性がある。一方で、IG の数値積分に関する分析の報告はなく、最適な積分ステップ数を導出する指針もない。そこで本研究では、積分誤差に関する分析を実施する。実験により、既存研究で広く用いられているステップ数の設定方法では、半数以上のインスタンスで過剰なステップ数となり、一部のインスタンスではステップ数が不足することを示す。本稿は IG のステップ数を分析した初めての研究報告であり、ステップ数を固定する既存の積分方法の問題点を明らかにする。

1　はじめに

特徴量帰属法 [1] とは、機械学習モデルが出力に寄与した特徴量を明らかにする方法であり、ブラックボックスである機械学習モデルの挙動を説明・分析することを目的に使用されている。特に特徴量帰属法の一種である **Integrated Gradients (IG)** [2] は、その数学的性質と低い計算コストから、画像処理分野 [3, 4] や言語処理分野 [5, 6, 7] で使用されている。

IG の持つ数学的性質の中でも重要視されるのが**完全性 (Completeness)** [2] である。特徴量の重要度である寄与度の加法性を保証する公理であり、寄与値を用いた演算や比較等の分析結果に正当性を担保する上で必要となる。

一方で、IG の計算では勾配の数値積分を使用するため、計算誤差が発生し得る。この計算誤差は加法性の破綻を引き起こすが、現在まで誤差に関する分析は報告されていない。また、計算誤差を低減するためには、数値積分の精度を決定する積分ステップ数（以下、ステップ数と略記）の設定が重要になるが、既存研究では経験則で固定した数値を使用することが一般的であり（表 1）、適切なステップ数に関する指針は現状明らかになっていない。また、近年では BERT [8] のような多くのパラメータを持つ言語モデルが登場したが、これらのモデルに関するステップ数の設定は知見が不足している。

表 1　文章分類で IG を使用する際に設定されるステップ数：既存研究ではステップ数が全インスタンスに対して任意の値で静的に固定される。

ステップ数	モデル	文献
50	CNN	[9] [10]
50, 250	DistilBERT、RoBERTa、BERT	[7]
10, 30, 100, 300	DistilBERT、RoBERTa、BERT	[11]
1000	Linear / Logistic regression	[12]
100, 1000	BERT、LSTM	[13]

そこで、本研究では、言語モデルを含む複数の実験条件で IG におけるステップ数と計算誤差に関して定量的に分析をおこなう（4.1 節）。その結果、半数以上のインスタンスについては想定よりも少ないステップ数で誤差が収束する一方で、一部のインスタンスに対しては非常に大きなステップ数を設定しても誤差が収束しない事例があることを示す。

また、非常に大きなステップ数を設定しても誤差が収束しないインスタンスに対して、定性的な分析を実施する（4.2 節）。分析により、積分過程で勾配が激しく変化する特徴量に誤差が発生しやすいことを示す。

本研究は、ステップ数による IG の変化を分析した初めての研究である。本研研究の実験結果から、既存の経験則による固定ステップ数の数値積分が非効率的かつ計算誤差を発生させる要因になっていることが示唆される。さらに、インスタンス毎に動的にステップ数を設定することで、効率的かつ計算誤差を低減できる可能性を示す。

2 Integrated Gradients

特徴量帰属法 [1] は特徴量の寄与値を計算する方法の総称であり、寄与値は各特徴量がモデル出力（予測）にどの程度有用だったかを表す。Integrated Gradients (IG) [2] は特徴量帰属法の中で主要な手法であり、望ましい公理と単純な計算方法から、画像処理 [3, 4] や言語処理 [5, 6, 7] で広く使われている。

入力ベクトル $\boldsymbol{x} \in \mathbb{R}^n$ における i 番目の特徴量 x_i の寄与度を計算する IG は以下の式で定義される。

$$\mathrm{IG}_i(\boldsymbol{x}) = (x_i - x_i') \int_{\alpha=0}^{1} \frac{\partial F}{\partial x_i}(\boldsymbol{x}' + \alpha(\boldsymbol{x} - \boldsymbol{x}'))d\alpha \quad (1)$$

ここで、F は微分可能な機械学習モデル、$\boldsymbol{x}'$ はベースラインベクトル (3.2 節参照) を表す。

α が $[0, 1]$ の値域で変化する計算過程において、ベースラインベクトルから入力ベクトルへの直線経路に沿って勾配の数値積分を行う。ここで、数値積分の際のサンプリング点は数値積分法とステップ数によって決定される。各サンプリング点で逆伝搬の計算を行う必要があるため、ステップ数に比例して計算コストが増大する．しかし、ステップ数の不足は大きな積分誤差を招く。そのため、誤差と計算コストはトレードオフの関係にあり、適切なステップ数を設定する必要がある [14]。

IG の完全性公理　各次元 i における IG の総和は、与えられた入力 $\boldsymbol{x}$ に対するモデル出力値からベースラインベクトル $\boldsymbol{x}'$ に対するモデル出力値を引くことで得られる。

$$\sum_{i=1}^{n} \mathrm{IG}_i(\boldsymbol{x}) = F(\boldsymbol{x}) - F(\boldsymbol{x}') \quad (2)$$

この性質は完全性 (Completeness) [2] といい、 IG が満たす数学的な公理の一つである。

積分ステップ数の設定における課題　完全性公理は寄与値の加法性を保証するため、寄与値を用いた演算や分析の正当性を担保するために重要である。しかしながら、実用上は数値積分による誤差が発生するため、この公理が崩れる場合がある。誤差を十分に抑制するためには、十分なステップ数を設定する必要がある。いくつかの既存研究では、誤差の抑制に必要なステップ数について言及している。CNN モデルを用いた文章分類では 20 から 300 ステップ、LSTM を用いた翻訳タスクでは 100 から 1000 ステップを使用している [2]。しかし、多くの既存研究でステップ数は、経験則に基づいて設定された値であり、定量的な分析によって設定された値ではない（表 1）。また、BERT [8] のような多くのパラメータを持つ言語モデルに必要なステップ数について分析した研究は存在しない。現状のままであれば、 適切なステップ数を設定できず誤差を十分に減少できない状況で IG を使用してしまう可能性があり、IG の結果の信頼性に疑問が生じる。そのため、ここでは完全性公理に基づいて算出される IG の理論値と数値積分によって算出される IG の実測値の誤差を測定し、ステップ数による寄与値の変化を定量的に分析する。

3 実験設定

現在の言語処理分野では、BERT のような事前学習モデルを、文章分類のような下流タスクにファインチューニングする手法が一般的である。そのため、ここでは複数のデータセットで事前学習モデルのファインチューニングを行い、指定のタスクを十分に解けるモデルを用意した。その後、ステップ数に基づく IG の変化を観測した。

3.1 IG の誤差指標

相対誤差 (Relative error; RE) [2] は、理論値に対する実測値の乖離を測定する指標である。各ステップ数における相対誤差は以下の式で計算する。

$$\mathrm{RE}(\boldsymbol{x}) = \left| \frac{\sum_i \widetilde{\mathrm{IG}}_i(\boldsymbol{x}) - (F(\boldsymbol{x}) - F(\boldsymbol{x}'))}{F(\boldsymbol{x}) - F(\boldsymbol{x}')} \right| \quad (3)$$

ここで、$\sum_i \widetilde{\mathrm{IG}}_i(\boldsymbol{x})$ は数値積分により求めた実測値の IG の和である。相対誤差は IG の理論和値と実測値の乖離を表す。

3.2 ベースラインベクトル

IG における適切なベースラインベクトルの設定は、今もなお議論されている [15, 16, 13]。本実験では、ベースラインベクトルはモデルにとって最小限の情報量を持つべきであるという考え方に基づき、最大エントロピーベースラインを使用した [16]。このベクトルは、テストデータセットにおいてモデル出力が最も一様に分布したベクトルである。

3.3 データセット・モデル

文章分類において広く使用されている AG News [17]、20 News [18]、SST2 [19] データセットを用いた。AG News と 20 News は、ニュース記事とカ

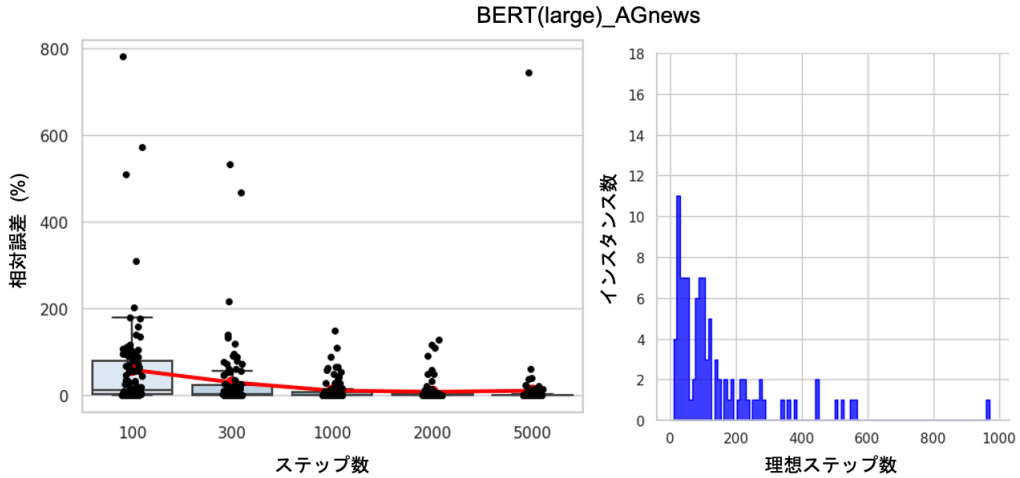


図 1　左の箱ひげ図：赤い線は各ステップ数の相対誤差平均を表し、各点は 1 インスタンスの相対誤差を表す。**右のヒストグラム**：各インスタンスの理想的なステップ数。縦軸はインスタンス数、横軸は理想ステップ数。インスタンスによって理想的なステップ数が異なるが、60 のインスタンスで理想ステップ数が 100 ステップ以内であった。

テゴリラベルから構成され、カテゴリ数はそれぞれ 4 つのカテゴリ、20 のカテゴリが存在する。SST2 は、映画のレビュー文と感情ラベルから構成され、肯定的・否定的の 2 つのラベルが存在する。データセットの詳細は付録 A.1 に示す。分類モデルに用いる言語モデルとして、BERT [8] と RoBERTa [20] のベースモデルとラージモデルの両方を使用した。モデルの詳細は付録 A.1 に示す。

3.4　その他の実験設定

IG を計算する際の数値積分法は右リーマン和とガウス・ルジャンドル積分を使用した。本研究の検証実験では各ステップ数における誤差を計算する必要があるため、非常に多くの実験時間を要する。現実的な実験時間に抑えるために、各データセットのテストデータから 100 個のインスタンスを無作為にサンプリングした。

4　実験結果

数値積分法として右リーマン和とガウス・ルジャンドル積分の 2 種類の方法を用いて実験を行ったが、右リーマン和がガウス・ルジャンドル積分よりも良い結果を示す傾向にあったため、ここでは右リーマン和の結果を報告する。詳細は付録 A.3 に示す。また、複数のモデルを用いて実験を行ったが、概ね同様の傾向を示したため、ここでは BERT(large)-AGnews を用いた結果のみを示し、他のモデルの結果は付録 A.4 に示す。

4.1　誤差の定量分析

指定したステップ数で IG を計算した場合に発生した誤差を定量的に解析した。ステップ数 100、300、1000、2000、5000 において IG を計算し、相対誤差を求めた。

図 1 左の箱ひげ図の結果から、膨大なステップ数であっても、大きな誤差を持つインスタンスが存在することが観測された。たとえ、ステップ数として 5000 を設定しても、相対誤差が 100%より大きいサンプルが存在し、最大で相対誤差が 700%より大きいインスタンスも確認した。これらの結果は、大きなステップ数を確保することが、より小さな誤差を保証するわけではないことを示す。

また、図 3 において、ステップ数の増加に伴い相対誤差が激しく上下動するインスタンスが観測された。この事例より、ステップ数の増加に伴い相対誤差が単調減少するわけではないことを確認できた。

より詳細な分析を行うため、各インスタンスにおける理想的なステップ数の分布を調べた。ここで、相対誤差が最初に 5%以内に収まるステップ数を理想的なステップ数として定義する [2]。図 1 右のヒストグラムから、1000 回以内のステップ数に理想的なステップ数が存在するインスタンスが 100 インスタンス中 98 インスタンス確認された。このうち、100 回のステップ数以内で理想的なステップ数を持つインスタンスが 60 存在した。この結果は、たとえ BERT のような言語モデルであっても、半数以上のインスタンスでは 100 回程度の少ないステップ数

ステップ数	相対誤差 (%)	可視化
1000	25%	Ġpeople Ġhave Ġbeen Ġkilled Ġin ĠKashmir Ġin Ġan Ġincrease
140	0.62%	Ġpeople Ġhave Ġbeen Ġkilled Ġin ĠKashmir Ġin Ġan Ġincrease
1000	17%	make sure the bike has cooled at least 6 hours since being run
130	0.94%	make sure the bike has cooled at least 6 hours since being run
1000	59%	russian oil giant si ##bn ##eft today rejected
520	4.8%	russian oil giant si ##bn ##eft today rejected

図 2　単語ごとの寄与：各行の上は固定ステップ数を想定し、下は理想ステップを使用した場合の可視化。

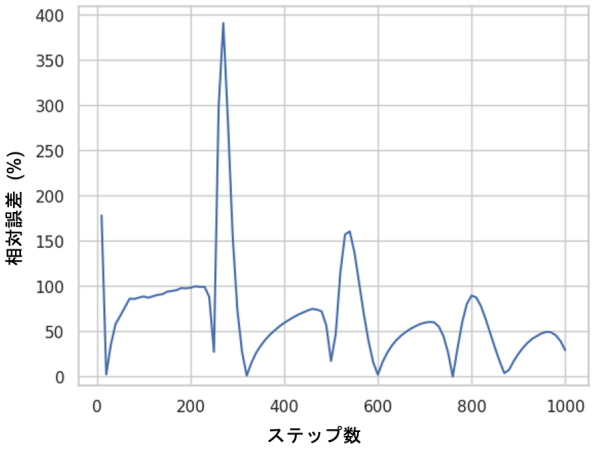


図 3　あるインスタンスのステップ数ごとの相対誤差：ステップ数を増やせば相対誤差が単調減少するわけではないことを示している。

でも問題なく誤差が収束することを示す。その一方で、1000 回以上のステップ数を設定したとしても、一部のインスタンスでは誤差は十分に収束しない。この実験によって、全インスタンスで固定のステップ数を指定する現在の IG の計算方法は、非効率的かつ完全性の保証が困難であることが示唆された。

4.2　誤差の定性分析

4.1 節の実験により、非常に大きなステップ数を設定した環境下でも、誤差が収束しないインスタンスが確認された。原因を探るため、誤差が収束しないインスタンスについて定性的な分析を行った。

図 2 に、誤差が発生するインスタンスに対する可視化の結果を示す。各単語に対応する各次元の寄与の和を各単語の寄与とし、緑色は正の寄与、赤色は負の寄与を表す。最も濃い色合いは、各単語寄与の中で最も大きな絶対値に割り当て、色は 0 に近づくにつれて薄くなる。このルールは文献 [2] で採用されている。誤差が発生した場合、全ての特徴量で寄与値が変化するのではなく、特定の特徴量に集中して値が大きく変化することが確認できる。この結果より、積分による IG の寄与値の集約の際に、入力ベクトルの全ての次元に関して小さな誤差が蓄積するのではなく、ある特定の次元に対して大きな誤差が発生すると推察される。

5　考察

分析から、インスタンス毎に必要なステップ数は異なることが明らかになった。したがって、特定の誤差を下回るまでステップ数を順次増やし、誤差を減らすことを推奨する。例えば、最初にステップ数を 2^n とし、誤差が一定に収束するまで n を増加させ実行することで、 IG が完全性を満たすことを保証できる。また、インスタンスごとにステップ数を最適化すれば、多くのインスタンスで必要なステップ数が削減できることが本実験の分析で示されているため、より効率的に IG を計算することができる。

6　結論

IG におけるステップ数は、データセットやモデルごとに研究者が経験則に基づいて主観的に決定しており、 IG の信頼性に疑問がある。本研究では、ステップ数ごとの誤差を定量的に分析した。その結果、理想的なステップ数はインスタンスごとに異なること、十分なステップ数を確保しても、誤差の減少を保証することはできないことを明らかにした。これらの結果から、モデルやデータセットごとにステップ数を固定する現在の主流の方法では、例え十分なステップ数を確保したとしても、結果が破綻したインスタンスが生成され、 IG の解析結果の信頼性が損なわれるリスクがあることがわかった。これを解決するために、インスタンスごとにステップ数を設定することを提案した。本研究は、ステップ数による IG 値の変動を定量的に分析し、既存の IG の問題点を明らかにした初めての研究である。

謝辞

本研究は JST ムーンショット型研究開発事業 JPMJMS2011-35（fundamental research) の支援を受けたものである。

参考文献

[1] Sheikh Rabiul Islam, William Eberle, Sheikh Khaled Ghafoor, and Mohiuddin Ahmed. Explainable artificial intelligence approaches: A survey. **ArXiv**, Vol. abs/2101.09429, , 2021.

[2] Mukund Sundararajan, Ankur Taly, and Qiqi Yan. Axiomatic attribution for deep networks, 2017.

[3] Julius Adebayo, Justin Gilmer, Michael Muelly, Ian Goodfellow, Moritz Hardt, and Been Kim. Sanity checks for saliency maps, 2020.

[4] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions, 2019.

[5] Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 10285–10299, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[6] Sandipan Sikdar, Parantapa Bhattacharya, and Kieran Heese. Integrated directional gradients: Feature interaction attribution for neural NLP models. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 865–878, Online, August 2021. Association for Computational Linguistics.

[7] Joseph Enguehard. Sequential integrated gradients: a simple but effective method for explaining language models, 2023.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding, 2019.

[9] Frederick Liu and Besim Avci. Incorporating priors with feature attribution on text classification, 2019.

[10] Lucas Dixon, John Li, Jeffrey Sorensen, Nithum Thain, and Lucy Vasserman. Measuring and mitigating unintended bias in text classification. In **Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society**, AIES '18, p. 67–73, New York, NY, USA, 2018. Association for Computing Machinery.

[11] Soumya Sanyal and Xiang Ren. Discretized integrated gradients for explaining language models, 2021.

[12] Tessa Han, Suraj Srinivas, and Himabindu Lakkaraju. Which explanation should i choose? a function approximation perspective to characterizing post hoc explanations, 2022.

[13] Jasmijn Bastings, Sebastian Ebert, Polina Zablotskaia, Anders Sandholm, and Katja Filippova. "will you find these shortcuts?" a protocol for evaluating the faithfulness of input salience methods for text classification, 2022.

[14] Gabriel Erion, Joseph D. Janizek, Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Improving performance of deep learning models with axiomatic attribution priors and expected gradients, 2020.

[15] Pascal Sturmfels, Scott Lundberg, and Su-In Lee. Visualizing the impact of feature attribution baselines. **Distill**, 2020. https://distill.pub/2020/attribution-baselines.

[16] Hanxiao Tan. Maximum entropy baseline for integrated gradients. In **2023 International Joint Conference on Neural Networks (IJCNN)**, pp. 1–8, 2023.

[17] Antonio Gulli. Agnews, 2004. http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html.

[18] Youngjoong Ko. A study of term weighting schemes using class information for text classification. In **Proceedings of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval**, SIGIR '12, p. 1029–1030, New York, NY, USA, 2012. Association for Computing Machinery.

[19] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In **Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing**, pp. 1631–1642, Seattle, Washington, USA, October 2013. Association for Computational Linguistics.

[20] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

A　付録

A.1　データセット・モデル

表 2 に本研究で使用したデータセットの詳細を示す。また表 3 に本研究で使用したモデルのテストデータにおける正解率を示す。

表 2　データセットの詳細

データセット	訓練 / テスト	クラス数	最大系列長
AG news	120k / 7.6k	4 クラス	50
20 news	11.3k / 7.53k	20 クラス	200
SST2	6.92k / 1.82k	2 クラス	20

表 3　各モデルのテストデータにおける正解率

モデル	正解率		
	AG News	20 News	SST-2
BERT-base (110M)	0.94	0.64	0.86
BERT-large (340M)	0.93	0.65	0.87
RoBERTa-base (125M)	0.94	0.61	0.88
RoBERTa-large (561M)	0.93	0.64	0.88

A.2　理想ステップ数

現実的な実験時間に抑えるために、理想ステップ数は 1000 回以内の範囲で測定している。表 4 に示すように、100 のインスタンス中、ほとんど全てのインスタンスの理想ステップ数が 1000 回以内に収まっている。

表 4　理想ステップ数が 1000 以内のインスタンス数

モデル	AG News	20 News	SST-2
BERT-base	100	100	100
BERT-large	98	99	99
RoBERTa-base	99	100	95
RoBERTa-large	99	100	97

A.3　2 つの積分方法の比較

右リーマン和とガウスルジャンドル積分の誤差低減においての比較を図 4 に報告する。紙面の都合上全ての実験設定における結果を載せることができないため、BERT_large モデルにおける実験設定の結果のみを報告する。図 4 において、右リーマン和の平均相対誤差は、ガウスルジャンドル積分のものと比べステップ数を増大するにつれて、より早く平均相対誤差が減少していることがわかる。報告していない実験設定においても同じ傾向が見られた。

A.4　誤差の定量分析

4.1 節において報告した誤差の定量分析について、その他の実験設定の結果も報告する。紙面の都合上全ての実験設定における結果を載せることができないため、図 5 に BERT (large) の結果とより特徴的な結果となった RoBERT (large)_AGnews についての結果を報告する。多くのステップ数を確保しても誤差が減少しないインスタンスが存在する。特に RoBERT (large)_AGnews については、ステップ数の増大が誤差低減に必ずしも繋がらないことを顕著に表している。また多くのインスタンスの理想ステップ数が 100 近くに存在する。報告していない実験設定においても同じ傾向が見られた。

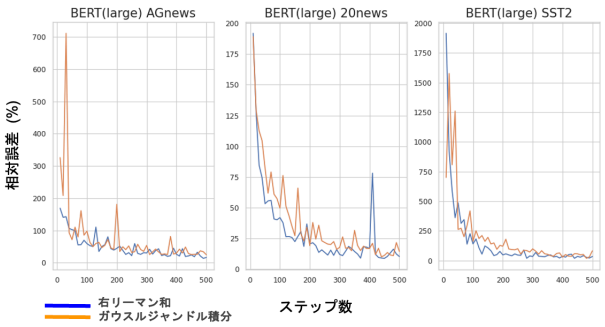
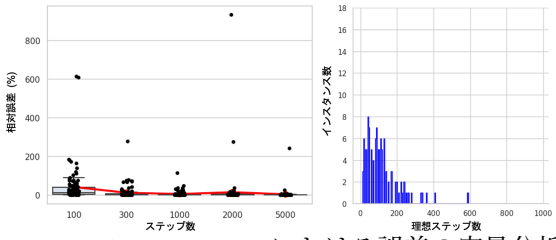
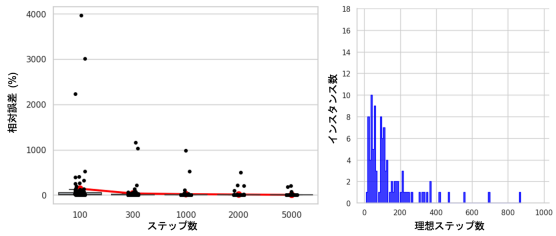


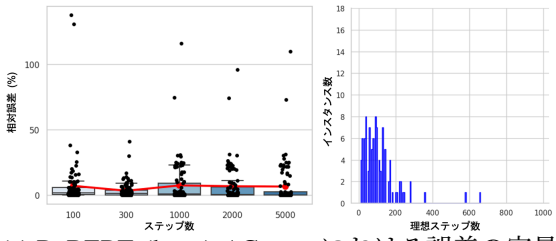
図 4　**右リーマン和とガウスルジャンドル積分それぞれの平均相対誤差**。青色が右リーマン和の結果で、オレンジ色がガウスルジャンドル積分の結果。



(a) BERT(large)_20News における誤差の定量分析



(b) BERT(large)_SST2 における誤差の定量分析



(c) RoBERTa(large)_AGnews における誤差の定量分析

図 5　**左の箱ひげ図**：赤い線は各ステップ数の相対誤差平均を表し、各点は 1 インスタンスの相対誤差を表す。**右のヒストグラム**：各インスタンスの理想的なステップ数。縦軸はインスタンス数、横軸は理想ステップ数。