

ステップニューロン集団としての 活性化関数解釈に基づく重み学習

佐村木 康太

金沢大学人間社会学域学校教育学類附属高等学校 1 年

概要

ニューラルネットワークにおける活性化関数は、一般には非線形性を導入するための工学的要素として説明されることが多い。一方で、sigmoid 関数は、閾値の異なる多数のステップニューロンからなる集団の平均発火率として解釈できる。本研究では、この考え方を出発点として、通常の 1 つのニューロンを多数のステップニューロンの集団として捉え直し、活性化関数を「参加ニューロン数」と「発火割合」の組合せとして説明する枠組みを考えた。さらに、既存のニューラルネットワークでは、1 つの活性化関数の背後にある小ニューロン群が暗に同じ重みを共有している点に着目し、小ニューロンごとの重みのばらつきを平均・分散・共分散によって近似する Moment モデルを提案した。共分散をそのまま学習するとパラメータ数が大きくなるため、実装では低ランク近似を用いた。MNIST および FashionMNIST による検証では、同等のパラメータ数を持つ通常の sigmoid ネットワークと比較して、提案モデルが高い精度と速い収束を示す傾向が確認された。

1 はじめに

ニューラルネットワークにおいて、活性化関数はほぼ必ず用いられる構成要素である。活性化関数の説明として最も一般的なのは、「非線形性を導入することで表現力を高める」という工学的な説明である。しかし、ニューラルネットワークがもともと生物の神経細胞を参考にして作られたモデルであることを考えると、活性化関数にも何らかの生物学的な説明が与えられてよいはずである。

本研究を始めるきっかけは、ED 法 (Error Diffusion Learning Algorithm) に関する研究を行う中で、生物学的に自然な学習を目指す枠組みであっても sigmoid 関数が使われていることに疑問を持ったことである。調査を進めると、Wilson と Cowan は、局所的なニューロン集団の活動を記述するモデルにおいて、集団応答関数として sigmoid 型関数を用いていた [1, pp. 3–4]。この sigmoid 型応答は単に仮定された非線形性ではなく、細胞ごとの閾値分布や求心性結合数の分布が単峰である場合に自然に現れるものとして説明されている [1, pp. 3–4]。また、Wilson と Cowan は、このような集団の入出力関係に関する先行研究として、Rall による単シナプス反射の理論的・実験的研究を挙げている [1, p. 4]。該当する Rall の研究は、理論的研究 [2] と実験的研究 [3] として発表されている。したがって本研究では、sigmoid 関数を単なる非線形関数ではなく、多数のステップニューロンの平均発火率を表す関数として捉える。

一方で、現在の深層学習では ReLU, GELU, SiLU など、多様な活性化関数が使われている [4, pp. 92–94]。これらは主に工学的な理由から導入されており、sigmoid と同様の生物学的解釈が十分に与えられているとは言いにくい。また、仮に sigmoid をステップニューロン集団の平均応答とし

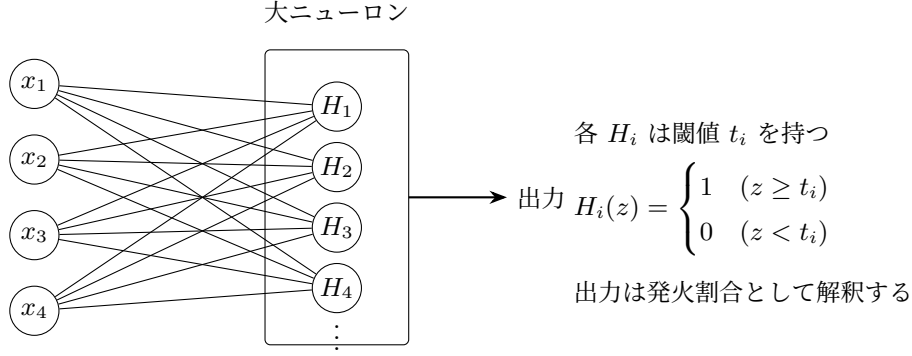


図1 1つのニューロンを多数のステップニューロン集団として捉える解釈

て捉えても、通常のニューラルネットワークでは、その背後にある小ニューロン群がすべて同じ重みを持つことになってしまう。これは、実際の神経細胞においてシナプス結合の強さが細胞ごとに異なることを考えると不自然である。

そこで本研究では、活性化関数をステップニューロン集団の応答として再解釈し、その内部にある小ニューロンごとの重みの違いを学習可能にするモデルを考えることを目的とした。

2 活性化関数のステップニューロン集団による解釈

2.1 ステップニューロン

まず、閾値 t を持つステップニューロンを次のように定義する。

$$H_t(x) = \begin{cases} 1 & (x \geq t), \\ 0 & (x < t). \end{cases} \quad (1)$$

これは、入力に閾値を超えたときに発火する単純なニューロンである。本研究では、通常の1つのニューロンを、このようなステップニューロンが多数集まった集団として捉える(図1)。

入力の総和を z とする。閾値 t が確率分布 ν に従って分布しているとき、入力 z に対して発火するステップニューロンの割合は

$$f(z) = \nu((-\infty, z]) \quad (2)$$

で表される。したがって、活性化関数 f は閾値分布 ν の累積分布関数として解釈できる。例えば、閾値分布がロジスティック分布であれば sigmoid 関数になり、正規分布であれば標準正規分布の累積分布関数 Φ になる。また、sigmoid は Φ の近似としても使うことができる。

2.2 参加ニューロン数と発火割合

上の解釈では、活性化関数は発火割合として説明される。しかし、ReLU や GELU のように出力の大きさが入力とともに増加し続ける関数は、発火割合だけでは説明しにくい。そこで本研究では、活性化関数を

$$a(z) = n(z)p(z) \quad (3)$$

という形で捉える。ここで、 $n(z)$ は入力 z に対して応答に参加するステップニューロン数、 $p(z)$ はそのうち発火する割合である。この見方を使うと、代表的な活性化関数は表1のように解釈できる。

表 1 ステップニューロン集団として見た活性化関数の例

活性化関数	参加ニューロン数 $n(z)$	発火割合 $p(z)$
Step	1	$H(z)$
Sigmoid	1	$\text{sigmoid}(z)$
ReLU	z	$H(z)$
GELU	z	$\Phi(z)$
SiLU	z	$\text{sigmoid}(z)$

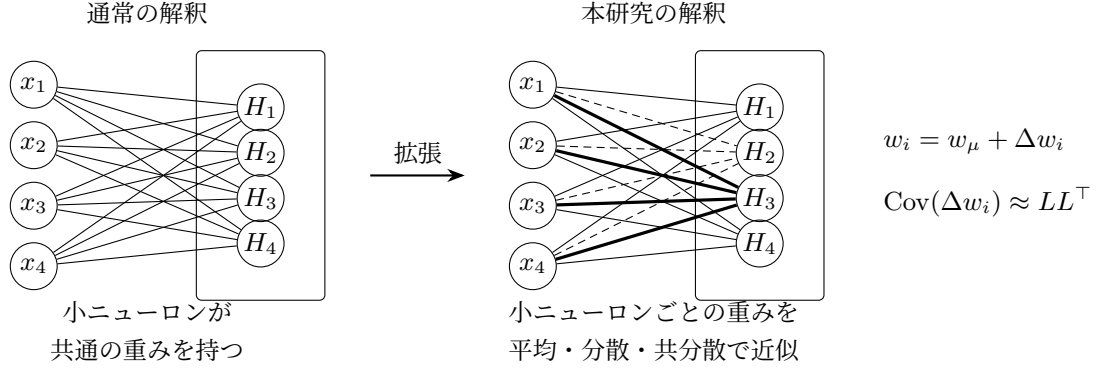


図 2 小ニューロンごとの重みのばらつきを導入する考え方

例えば ReLU は $zH(z)$ と書けるので、入力 z に比例して参加ニューロン数が増え、それらが閾値 0 のステップ関数で発火する場合として説明できる。GELU は $z\Phi(z)$ と書けるため、参加ニューロン数が z に比例し、発火割合が正規分布の累積分布関数で与えられる場合として解釈できる。このように考えると、複数の活性化関数を「参加ニューロン数」と「発火割合」という共通の枠組みで整理できる。

3 既存モデルの問題点

ステップニューロン集団として活性化関数を解釈すると、既存のニューラルネットワークには次の問題がある。通常のネットワークでは、あるニューロンの入力

$$z = w^\top x + b \quad (4)$$

と書かれ、その後に $\text{sigmoid}(z)$ や $\text{ReLU}(z)$ などの活性化関数が適用される。しかし、活性化関数の背後に多数のステップニューロンがあると考えれば、それぞれの小ニューロンは本来、異なる重みを持っていてもよいはずである。

ところが、通常の表現では、すべての小ニューロンが同じ w を共有していることになる。つまり、既存の活性化関数は、小ニューロン集団を考えているにもかかわらず、その内部の重み自由度を消してしまっていると見ることができる。そこで本研究では、この小ニューロンごとの重みのばらつきを明示的に導入し、それを効率よく近似して学習するモデルを考える（図 2）。

4 Moment モデル

4.1 集団平均発火率

大ニューロンの内部に m 個の小ニューロンがあるとする。 i 番目の小ニューロンの重みを $w_i \in \mathbb{R}^d$, 閾値を T_i とすると, その発火は

$$s_i(x) = H(w_i^\top x - T_i) \quad (5)$$

で表される。大ニューロンの出力を小ニューロンの平均発火率とすると,

$$y_m(x) = \frac{1}{m} \sum_{i=1}^m s_i(x) \quad (6)$$

である。小ニューロン数 m が十分大きいとき, 経験平均は期待値に近づくため,

$$y(x) = \mathbb{E} [H(w_i^\top x - T_i)] = P(w_i^\top x - T_i > 0) \quad (7)$$

と近似できる。

4.2 重みのばらつきとモーメント近似

小ニューロンごとの重みと閾値を, 平均成分と個体差に分けて

$$w_i = w_\mu + \Delta w_i, \quad (8)$$

$$T_i = T_\mu + \Delta T_i \quad (9)$$

と書く。このとき,

$$w_i^\top x - T_i = w_\mu^\top x - T_\mu + (\Delta w_i^\top x - \Delta T_i) \quad (10)$$

である。ここで

$$\mu = w_\mu^\top x - T_\mu \quad (11)$$

とおき, $\Delta w_i^\top x - \Delta T_i$ を平均 0 の確率変数とみなす。重みのずれの共分散を

$$\Sigma = \text{Cov}(\Delta w_i) \quad (12)$$

とすると,

$$\text{Var}(\Delta w_i^\top x) = x^\top \Sigma x \quad (13)$$

である。閾値のばらつきやその他の内部ノイズをまとめて 1 と規格化すると, 分散は

$$x^\top \Sigma x + 1 \quad (14)$$

と書ける。したがって, 集団平均発火率は標準正規分布の累積分布関数 Φ を用いて

$$y(x) \approx \Phi \left(\frac{\mu}{\sqrt{x^\top \Sigma x + 1}} \right) \quad (15)$$

と近似できる。

この式は, 通常の sigmoid 型の活性化関数に対して, 小ニューロン内部の重みのばらつきを反映したものと見ることができる。分母が入力方向に依存するため, 入力 x が重み分布のどの方向に沿っているかによって発火の鋭さが変化する。

4.3 低ランク近似

入力次元が大きい場合、例えば MNIST では $d = 784$ であるため、 $d \times d$ の共分散行列 Σ をそのまま学習するのは計算量とメモリの両面で現実的でない。そこで、 $r \ll d$ として

$$\Sigma \approx LL^\top, \quad L \in \mathbb{R}^{d \times r} \quad (16)$$

と低ランク近似する。このとき、

$$x^\top \Sigma x \approx x^\top LL^\top x = \|L^\top x\|_2^2 \quad (17)$$

である。さらに、共分散の近似誤差や入力と独立な内部ノイズを吸収するための項 s_b^2 を加えると、実装上の Moment sigmoid は

$$\text{MomentSigmoid}(x) = \Phi \left(\frac{w_\mu^\top x + b}{\sqrt{\|L^\top x\|_2^2 + s_b^2 + 1}} \right) \quad (18)$$

となる。ここで s_b は正になるように、実装では softplus などを通して学習する。

このモデルでは、小ニューロンをすべて明示的に持つのではなく、その集団の統計的な挙動を平均と分散で近似する。したがって、多数の小ニューロンを直接実装するよりも少ないパラメータで、内部の重み自由度をある程度表現できる。

5 実験

5.1 比較するモデル

実験では、通常の sigmoid ネットワークを Base モデル、提案する Moment sigmoid を用いたネットワークを Moment モデルと呼ぶ。MNIST および FashionMNIST を用いて、両モデルのパラメータ数がほぼ同じになるように比較した。共通の実験設定を表 2 に示す。

表 2 実験設定

項目	設定
データセット	MNIST, FashionMNIST
シード数	20
エポック数	30
ネットワーク	中間層 1 層の MLP
Base モデル	sigmoid 関数, 256 ユニット, 203,550 params
Moment モデル	Moment sigmoid, 251 ユニット, 203,842 params
低ランク次元	$r = 16$

Base モデルと Moment モデルで隠れ層のユニット数が異なるのは、パラメータ数をできるだけ揃えるためである。Moment モデルは、通常の重みに加えて低ランク行列や内部ノイズのパラメータを持つため、同じユニット数では Base モデルよりパラメータ数が多くなる。

5.2 結果

図 3, 図 4 に, 20 シードに対する学習曲線を示す. 線はシード方向の中央値, 帯は四分位範囲 (IQR) である. また, 30 エポック終了時点の平均値と標準偏差を表 3 に示す. 表中の太字は, 各データセット・各指標で良い平均値を示す. Test Acc は大きいほど良く, Test NLL と Test ECE は小さいほど良い.

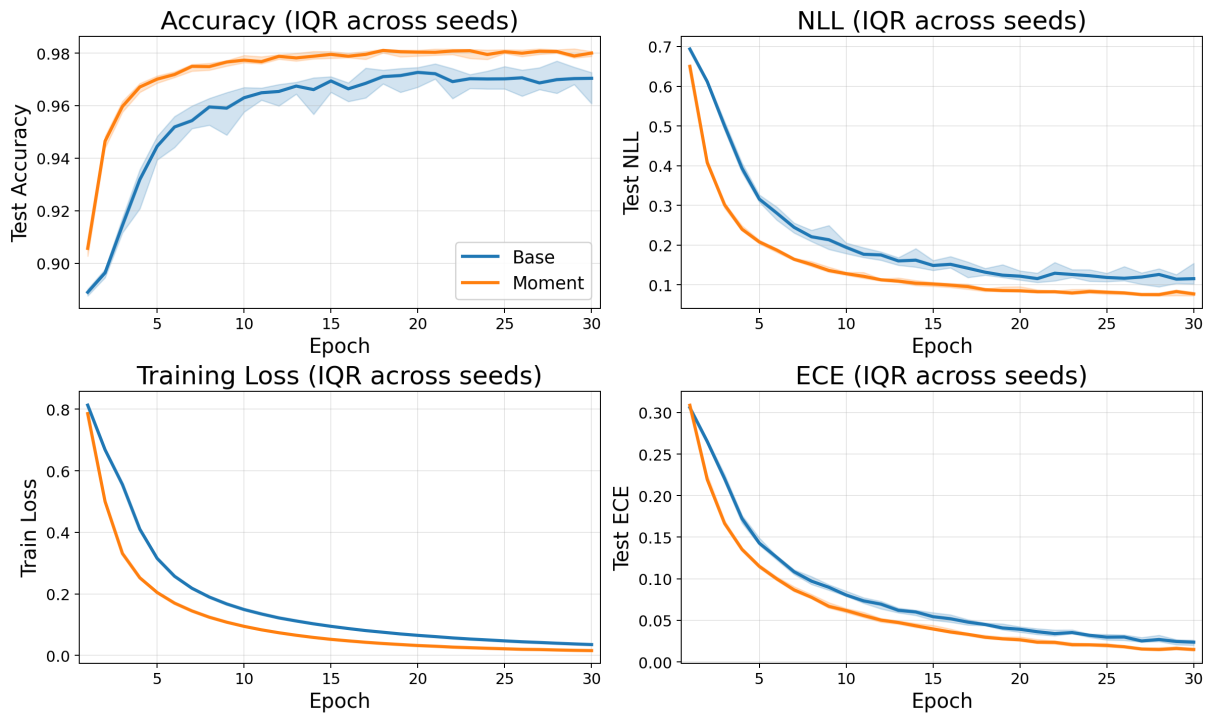


図 3 MNIST における Base モデルと Moment モデルの比較. 線は中央値, 帯は IQR を表す.

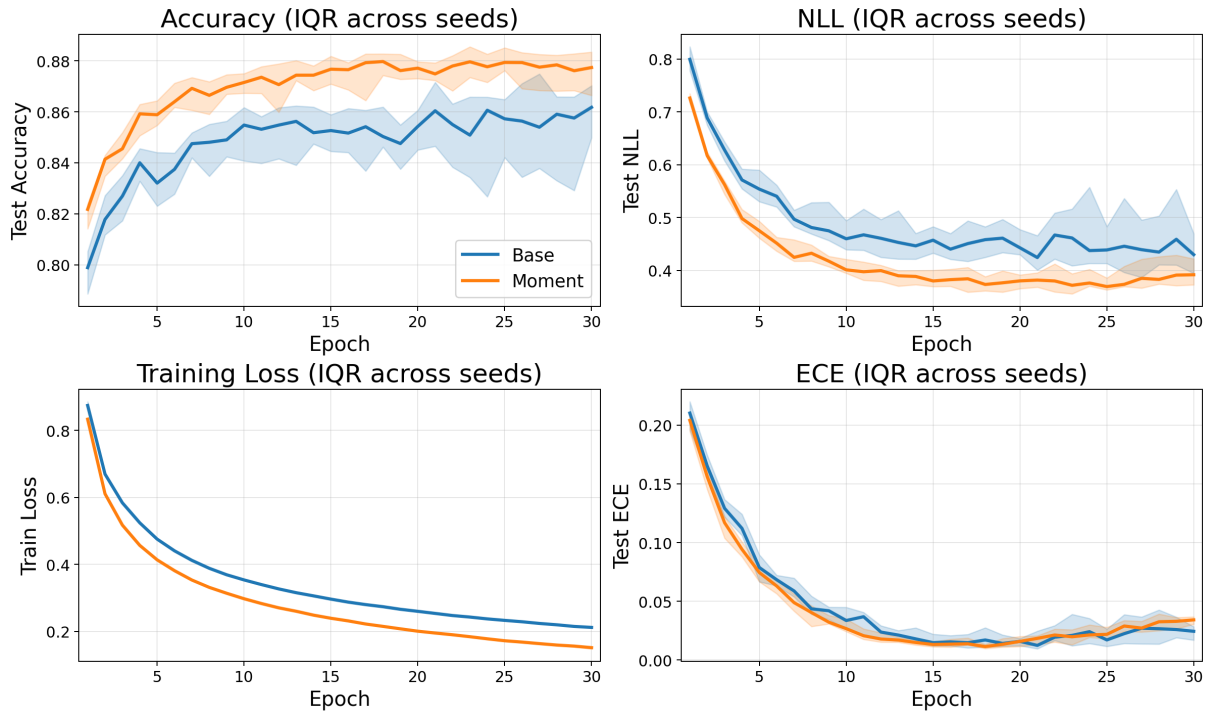


図4 FashionMNIST における Base モデルと Moment モデルの比較. 線は中央値, 帯は IQR を表す.

表3 30 エポック終了時点の結果 (平均 $\pm$ 標準偏差). 太字は良い値を表す.

データセット	モデル	Test Acc	Test NLL	Test ECE
MNIST	Base	0.9665 $\pm$ 0.0097	0.1268 $\pm$ 0.0329	0.0243 $\pm$ 0.0051
MNIST	Moment	0.9790 $\pm$ 0.0038	0.0810 $\pm$ 0.0152	0.0148 $\pm$ 0.0029
FashionMNIST	Base	0.8578 $\pm$ 0.0170	0.4440 $\pm$ 0.0648	0.0281 $\pm$ 0.0178
FashionMNIST	Moment	0.8761 $\pm$ 0.0092	0.3975 $\pm$ 0.0334	0.0346 $\pm$ 0.0062

MNIST では, Moment モデルは Base モデルと比較して, 初期のエポックから test accuracy が高く, test NLL, training loss, ECE もより速く低下した. 30 エポック終了時点でも, test accuracy, test NLL, test ECE のいずれも Moment モデルの方が良い平均値を示した.

FashionMNIST でも, Moment モデルは test accuracy, test NLL, training loss において Base モデルより良い傾向を示した. 一方で, 30 エポック終了時点の test ECE だけを見ると, 平均値は Base モデルの方が小さかった. したがって, 今回の結果から強く言えるのは, 提案モデルが精度と収束速度を改善する傾向を示したという点であり, 較正誤差についてはデータセットや学習後半の挙動を含めて追加検証が必要である.

6 考察

本研究の結果から, 活性化関数を単なる非線形関数ではなく, ステップニューロン集団の統計的応答として捉えることが, モデル設計に有効な視点を与える可能性が示された. 特に, 通常の sigmoid ネットワークでは無視されていた小ニューロンごとの重みのばらつきを, 平均・分散・共分散の近似

によって学習可能にしたことで、学習初期の収束が速くなったと考えられる。

一方で、最終精度の差が常に大きいとは限らない。また、FashionMNIST の ECE のように、すべての指標で一貫して Moment モデルが優れるわけではない可能性もある。この点から、Moment モデルの効果は単純に「常に性能を上げる」というよりも、入力方向に依存した分散項を導入することで最適化や表現の性質を変えるものだと考えられる。低ランク次元 r も大きければ大きいほどよいわけではなく、タスクごとに適切な値が存在すると考えられる。

本研究のもう 1 つの意義は、sigmoid だけでなく、ReLU, GELU, SiLU などの活性化関数も、「参加ニューロン数」と「発火割合」という同じ枠組みで説明できる可能性を示した点である。今後、この考え方を使って、既存の活性化関数の比較や、新しい活性化関数の設計を行える可能性がある。

7 今後の課題

今後の課題は主に 4 つある。

第 1 に、MNIST や FashionMNIST 以外のタスクで有効性を検証する必要がある。今回の実験では画像分類の比較的基本的なデータセットを用いたため、より複雑なデータセットや異なる種類のタスクで同様の傾向が見られるかを調べたい。

第 2 に、ReLU や GELU など、sigmoid 以外の活性化関数に対して Moment 化を行う必要がある。本研究の枠組みでは、発火割合 $p(z)$ の部分を Moment 化すれば、他の活性化関数にも拡張できると考えられる。

第 3 に、低ランク次元 r やグループ数などのハイパーパラメータの影響を詳しく調べる必要がある。特に、 r の自動最適化ができれば、モデルをより実用的にできる可能性がある。

第 4 に、BatchNorm や Dropout など、ニューラルネットワークで広く使われている技術を同じ枠組みで説明できるかを調べたい。これらの技術も、神経細胞集団の統計的性質として解釈できれば、生物学的な発想から新しい構造を設計することにつながると考えられる。

8 まとめ

本研究では、通常の 1 つのニューロンを多数のステップニューロンの集団として捉え直し、活性化関数を「参加ニューロン数」と「発火割合」の組合せとして説明する枠組みを提案した。また、既存の活性化関数では、小ニューロン集団が暗に同じ重みを共有しているという問題に着目し、小ニューロンごとの重みのばらつきを平均・分散・共分散で近似して学習する Moment モデルを構築した。

MNIST および FashionMNIST による実験では、同等パラメータ数の通常の sigmoid ネットワークと比較して、提案モデルが精度と収束速度の面で良い結果を示す傾向が確認された。この結果は、活性化関数を生物学的な集団応答として捉え直すことが、ニューラルネットワークの設計に新しい視点を与える可能性を示している。

参考文献

- [1] H. R. Wilson and J. D. Cowan, “Excitatory and inhibitory interactions in localized populations of model neurons,” *Biophysical Journal*, vol. 12, no. 1, pp. 1–24, 1972.

- [2] W. Rall, “A statistical theory of monosynaptic input-output relations,” *Journal of Cellular and Comparative Physiology*, vol. 46, no. 3, pp. 373–411, 1955.
- [3] W. Rall, “Experimental monosynaptic input-output relations in the mammalian spinal cord,” *Journal of Cellular and Comparative Physiology*, vol. 46, no. 3, pp. 413–437, 1955.
- [4] S. R. Dubey, S. K. Singh, and B. B. Chaudhuri, “Activation functions in deep learning: A comprehensive survey and benchmark,” *Neurocomputing*, vol. 503, pp. 92–108, 2022.