6M-02

# 損失関数を基礎とした事前分布をもつベイジアンニューラルネットワーク

葛原 優樹 †　浦田 光佑 ‡　安田 宗樹 §
山形大学大学院理工学研究科 †　山形大学工学部 ‡　山形大学大学院理工学研究科 §

## 1. はじめに

ディープニューラルネットワーク（deep neural network（DNN））は順伝播型ニューラルネットワークを多段にした数理モデルであり，深層学習の一種として広く使用されている [1]．DNN の学習において，同一データに対して異なる初期値で複数回学習を行い得られた DNN のアンサンブルで予測するといった方法が取られることがある．このアンサンブル法は，単一モデルの予測と比べ安定的に良好な予測結果を与えることが経験的に知られている．この方法は，パラメータの事前分布を損失関数に対するボルツマン分布とした場合のベイジアンニューラルネットワーク（Bayesian neural network（BNN））と見なすことができる．本講演では，このアンサンブル法を BNN として再解釈し，その理論背景と利点について議論を行う．

## 2. ベイジアンニューラルネットワーク

DNN の通常の学習は，損失関数の最小化により学習パラメータを最適化することである（点推定）．それに対し，BNN は具体的なパラメータの値ではなく，それらの分布を最適化する手法である（分布推定） [2]．

入力データ $\boldsymbol{x}$ に対して，クラス $C$ を出力する DNN があるとする．学習パラメータを $\theta$ とすると，DNN は $P(C \mid \boldsymbol{x};\theta)$ で表される．この DNN のパラメータの分布の最適化は，

$$P(C \mid \boldsymbol{x}) = \int P(C \mid \boldsymbol{x};\theta) P_{\mathrm{pri}}(\theta) d\theta \tag{1}$$

によって行う．$P_{\mathrm{pri}}(\theta)$ は事前分布であり，$\theta$ に関する事前の仮説を表している．BNN においてこの事前分布をうまく設計することが非常に重要となる．本来の BNN ではこの事前分布をデータからの学習で決定することになるのだが，本研究では，次節の考え方に基づいて事前分布をモデル化し，そのモデル化された事前分布の下での BNN を考える．

## 3. 事前分布のモデル化

通常の DNN の学習は，損失関数 $\ell(\theta)$ の最小化によってなされる．しかし，損失関数を最小化する $\theta$ は一般に過学習の可能性があり，最適なパラメータとは限らない．最適な $\theta$ は損失関数の最小点 $\theta^*$ からズレていると考えるのが自然である．そこで，そのズレを考慮に入れるため，損失関数の最小点を中心にいくらかの幅を持ったような分布を事前分布 $P_{\mathrm{pri}}(\theta)$ として採用したい．

その目的のために，変分自由エネルギー

$$\mathcal{F}[\mathcal{Q}] = \int \ell(\theta)\mathcal{Q}(\theta)d\theta + \frac{1}{\beta}\int \mathcal{Q}(\theta)\ln \mathcal{Q}(\theta)d\theta \tag{2}$$

を最小化する分布 $\mathcal{Q}(\theta)$ を事前分布 $P_{\mathrm{pri}}(\theta)$ とする．変分自由エネルギーの第 1 項は DNN の損失関数最小化を表しており，第 2 項は分布に幅を持たせるためのエントロピー項を表している．逆温度 $\beta > 0$ が幅の規模を調整する．式 (2) の最小化は変分法により実行され，結果，ボルツマン分布

$$\mathcal{Q}^*(\theta \mid \beta) = \frac{1}{Z(\beta)} \exp\left(-\beta\ell(\theta)\right) \tag{3}$$

を得る．ここで，$Z(\beta)$ は規格化定数である．$\beta \to \infty$ の極限（低温極限）ではエントロピーの効果が消失するため損失関数を最小とする $\theta^*$ のみに確率をもつデルタ関数 $\mathcal{Q}^*(\theta \mid \infty) = \delta(\theta - \theta^*)$ となる．この場合，式 (1) より，$P(C \mid \boldsymbol{x}) = P(C \mid \boldsymbol{x};\theta^*)$ となり，通常の損失関数最小化学習と等しくなる．一方，$\beta < \infty$ では，エントロピーの効果で $\mathcal{Q}^*(\theta \mid \beta)$ は $\theta^*$ を中心にいくらかの幅をもつ分布となる．議論の簡単化のため，以下の節では逆温度 $\beta$ は 1 に固定する．

## 4. BNN による推定とその背景

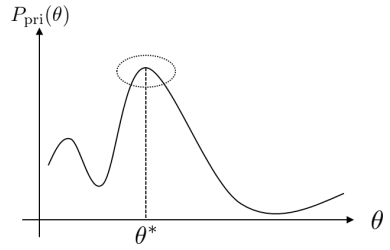


図 1　$P_{\mathrm{pri}}(\theta)$ のイメージ．

BNN を用いて推定を行うためには式 (1) によって行うことになるが，この積分は計算量が非常に大きくなるために計算することができない．そこでモンテカルロ積分による近似計算を行う．この時，式 (3) より，式 (1) の $P_{\mathrm{pri}}(\theta)$ は

$$P_{\mathrm{pri}}(\theta) = \frac{1}{Z}\exp(-\ell(\theta)) \tag{4}$$

であり，この分布からサンプル点を生成することにする．この際，図 1 の $\theta^*$ のように $P_{\mathrm{pri}}(\theta)$ を最大にする $\theta$ がもっともよいパラメータとは限らないがその周辺に存在していると考えられ，この方法によりそれらの点をサンプリングできると考えられる．サンプリング法は，DNN の通常の学習を複数回独立に実行し，その学習解を $P_{\mathrm{pri}}(\theta)$ からのサンプリング点と仮定

Bayesian Neural Network with loss-function-based prior

† Yuki Kuzuhara; Graduate School of Science and Enginnering, Yamagata University
‡ Kosuke Urata; faculty of Enginnering, Yamagata University
§ Muneki Yasuda; Graduate School of Science and Enginnering, Yamagata University

するという方法である．このようにサンプリングを行う背景には，DNN の確率的勾配降下法（stochastic gradient descent（SGD））とランジュバン・モンテカルロ法（Langevin Monte Carlo method（LMC））の類似がある．

DNN の学習は損失関数を微分することにより勾配を求め，それに従ってパラメータ更新を行う．すべての訓練データを使った場合（最急降下法（gradient descent（GD））の更新式は

$$\theta \leftarrow \theta - \varepsilon \frac{\partial L(\theta)}{\partial \theta} \tag{5}$$

で表される．ここで，$L(\theta)$ は損失関数を，$\varepsilon$ は学習率を表す．それに対して，分割したミニバッチごとに更新を行う方法が SGD である．ミニバッチ $B_1$，$B_2$，$\cdots$，$B_r$，$\cdots$，$B_R$ に対応した損失関数 $L_1(\theta)$，$L_2(\theta)$，$\cdots$，$L_r(\theta)$，$\cdots$，$L_R(\theta)$ があるとすると，SGD の更新式は

$$\theta \leftarrow \theta - \varepsilon \frac{\partial L_r(\theta)}{\partial \theta} = \theta - \varepsilon \frac{\partial L(\theta)}{\partial \theta} + (\text{ノイズ}) \tag{6}$$

となる．このように SGD と GD の間には誤差が生じ，これはノイズとして考えられる．

LMC はある関数 $U(\theta)$ について，

$$P_{\mathrm{LMC}}(\theta) = \frac{1}{Z_{\mathrm{LMC}}} \exp(-U(\theta)) \tag{7}$$

に従うサンプル点 $\theta$ を得るために使われる．ここで，$Z_{\mathrm{LMC}}$ は規格化定数である．LMC のサンプリングダイナミクスは

$$\theta \leftarrow \theta - \varepsilon \frac{\partial U(\theta)}{\partial \theta} + (\text{ノイズ}) \tag{8}$$

の通りである．

式 (6) と式 (8) が類似していることから，SGD を用いる方法で，LMC と同様にサンプリングを行うことができると考えられる．$K$ 回の通常の SGD 学習により，$\theta$ の学習解 $\theta_1, \theta_2, \cdots, \theta_K$ が集められたとする．そして，

$$P(C \mid \boldsymbol{x}) \approx \frac{1}{K} \sum_{k=1}^{K} P(C \mid \boldsymbol{x}; \theta_k) \tag{9}$$

のように $K$ 個の解を事前分布からのサンプリング点と仮定し，モンテカルロ積分により近似する．以上の一連の手続きは，複数回学習を行い得られた DNN のアンサンブルで予測することに対応している．

## 5. 数値実験

MNIST（手書き数字画像）と Fashion-MNIST（衣類画像）のデータセットを使用して数値実験を行った．どちらも $28 \times 28$ ピクセルの画像であり，10 種類のラベルが存在する．入力データの各画素値は 0–255 であるが，あらかじめ 255 で割り 0–1 の間に収まるように正規化を行う． 4 層の DNN を使用し，中間層の素子数は 2 層とも 500 で，活性化関数には ReLU を使う．学習に使用する勾配法は Adam [3] とし，損失関数は Cross Entoropy である．バッチサイズは 128，エポック数は 50 で行う．BNN のサンプル数は $K = 10$ とする．さらに，人工的にノイズを加えた場合についても頑健性を確かめるために，入力データの各画素値に対して独立に平均 0，分散 1 のガウスノイズを加える実験も行う．

まず，データセットをそのまま使用した場合の結果を表 1 に示す．MNIST，Fashion-MNIST どちらも DNN と比べ BNN の認識率が高くなっている．

表 1　ノイズなしの場合の結果

| | MNIST | Fashion-MNIST |
|---|---|---|
| DNN | 97.9% | 87.3% |
| BNN | 98.7% | 89.9% |

次に，訓練データに対してノイズを加え同様の実験を行う．その結果は表 2 に示す．これについても MNIST，Fashion-MNIST どちらも DNN と比べ BNN の認識率が高くなっている．

表 2　訓練データにノイズを加えた場合の結果

| | MNIST | Fashion-MNIST |
|---|---|---|
| DNN | 96.8% | 94.5% |
| BNN | 97.9% | 96.7% |

さらに，訓練データだけでなくテストデータに対してもノイズを加え同様の実験を行う．その結果を図 3 に示す．これについても MNIST，Fashion-MNIST どちらも DNN と比べ BNN の認識率が高くなっている．

表 3　訓練データとテストデータ両方にノイズを加えた場合

| | MNIST | Fashion-MNIST |
|---|---|---|
| DNN | 94.4% | 79.5% |
| BNN | 97.9% | 85.3% |

## 6. まとめ

本研究では，DNN のアンサンブル法に対して BNN という再解釈を行い，その理論背景について考察し，その効果を確かめた．その結果，データ中に含まれるノイズに対して頑健になっていることが分かった．Adam 以外の勾配法についても今後検討を行いたい．また，モンテカルロ積分のサンプリング点数を増やした場合についても性能検討を行うことが今後の課題である．

### 謝辞

本研究は科研費（18K11459，18H03303, 21K11778）及び JST CREST（JPMJCR1402）の助成を受けたものである．

### 文献

[1] 麻生 英樹・安田 宗樹・前田 新一・岡野原 大輔・岡谷 貴之・久保 陽太郎・ボレガラ ダヌシカ: 深層学習，近代科学社，2015.
[2] R. Neal: Bayesian learning for neural networks. Vol. 118. Springer Science & Business Media, 2012.
[3] Diederik and Jimmy: ADAM: A METHOD FOR STOCHASTIC OPTIMIZATION , 2014.