CF-006

Mixup はヘッシアン正則化を含むか？
Theoretical Insights into Mixup: Does It Include Hessian Regularization?

杉山 孔亮 [*1]　内田 真人 [*1]
Kosuke Sugiyama　Masato Uchida

概要

Mixup は，サンプルを凸結合することで新たなサンプルを生成するデータ拡張手法である．この手法は非常にシンプルでありながら，Deep Neural Network の汎化性能の向上に大きく貢献してきた．Mixup を理解するためにさまざまな研究が行われており，Mixup がヤコビアン正則化を含んでいることが理論的に示されている．さらに，Mixup と label smoothing の関連性が示唆されている．しかし，重要な点として，これらの研究は Mixup で生成されるサンプルが元のサンプルに近い場合に限定されており，他の場合については未だに解析が行われていない．本論文では，これまで解析されてこなかった Mixup で生成されるサンプルが元のサンプルから遠い場合において，Mixup をロジスティック回帰に適用することでヘッシアン正則化の効果が得られることを証明する．我々の理論的な結果により，Mixup は複数の正則化手法を組み合わせて近似する手法であるという新たな解釈が得られる．

1 序論

Mixup とは，サンプルを凸結合することで新たなサンプルを生成するデータ拡張手法である [1]．この手法は非常にシンプルな方法でありながら，多くの問題において Deep Neural Network（DNN）の汎化性能の向上に寄与することが確認されている [1]．これは，Mixup により生成されたサンプルが，元のサンプルの特徴を引き継ぎながらも，わずかに異なる特徴を有することが，DNN の学習において有用であるためだと考えられる．Mixup は，テーブルデータや画像の分類問題に限らず，音声認識 [2] やドメイン適応 [3]，グラフ分類 [4] など，様々な分野に応用されている．

Mixup が汎化性能の向上に寄与する理由については，Mixup により生成されたサンプルが元のサンプルの近傍にある場合における解析が行われている．これらの解析には，一般的に訓練サンプルの周りでのテイラー展開が用いられる．[5] では，テイラー展開を用いることで，Mixup を用いたときの経験損失がヤコビアン正則化を含むことを明らかにしている．さらに，[6] では，ヤコビアン正則化が入力に対する摂動の付与と等価であることが示されている．また，サンプルの出力情報であるラベルに対して一様なノイズを加える label smoothing [7] との関連性もさまざまな研究により示唆されている [8, 9, 10, 11]．実際，Mixup と label smoothing はいずれも，期待較正誤差（ECE）の改善と予測確率のエントロピーの増加という効果を持つことが経験的に示されている [8, 9, 10, 11]．

一方，Mixup によって生成されたサンプルが元のサンプルの近傍にない場合における解析は進んでいない．この理由の一つに，Mixup を使用することが必ずしも汎化性能の向上につながるとは限らず，一般的な解析を行うのが困難であることがある．例えば，Mixup によって生成されたサンプルが元のサンプルの近傍にない場合，Mixup で生成されたサンプルのラベルが，元の訓練データのラベルと一致しない場合がある．この現象は「manifold intrusion」として知られている [12, 13]．しかし，manifold intrusion が生じない理想的な状況であっても，Mixup が汎化性能の向上に寄与する理由は明らかにされていない．これは，訓練サンプルの周りでのテイラー展開を用いた解析が行えないということが技術的な障壁になっているためである．本論文では，Mixup によって生成されたサンプルが元のサンプルの近傍に無い場合について，manifold intrusion が生じない状況でのロジスティック回帰において，ヘッシアン正則化の効果があることを明らかにする．

この事実が成り立つことは，Mixup が本来的には訓練サンプル間での学習モデルの振る舞いを線形化することを意図した手法 [1] として提案されたものであることから類推することができる．すなわち，ある区間において学習モデルの振る舞いが線形であることは，その区間における学習モデルのヘッセ行列のすべての要素が 0 であることと等価であり，この事実が成り立つことは自然である．本論文では，この結果をテイラー展開を用いずに導いた．以上より，先行研究で示された結果に加え，本論文において新たに示された結果を踏まえると，ヤコビアン正則化，label smoothing，ヘッシアン正則化という複数の正則化手法をまとめて近似する手法として Mixup を特徴づけることができる．

本論文の構成は以下の通りである．**2** 節では，Mixup の定義と，既に行われている Mixup に関する理論研究について説明する．**3** 節では，Mixup がヘッシアン正則化を含むことについて，問題設定と解析結果を説明する．加えて，解析した学習モデルよりも複雑な学習モデルについても，解析結果と同様の傾向があることを数値実験により確かめる．**4** 節は，本論文のまとめである．

2 関連研究

2.1 Mixup の定式化

D 次元の入力空間を $\mathcal{X} \subseteq \mathbb{R}^D$，総ラベル数を L とし，出力空間を $\mathcal{Y} = [0,1]^L$ とする．入力点を $\boldsymbol{x} \in \mathcal{X}$，入力点に対応するラベルを $\boldsymbol{y} \in \mathcal{Y}$ とする．各サンプル $(\boldsymbol{x}, \boldsymbol{y})$ は真の分布 P_* に *i.i.d.* で従うとする．n 個の訓練データ $S = \{(\boldsymbol{x}_1, \boldsymbol{y}_1), \ldots, (\boldsymbol{x}_n, \boldsymbol{y}_n)\}$ が得られているとする．また，S の経験分布を $\widehat{P}_S$ と書く．学習に用いる損失関数を $l(\boldsymbol{y}, \widehat{\boldsymbol{y}})$ と書く．ただし，$\boldsymbol{y}$ は真のラベルで，$\widehat{\boldsymbol{y}}$ は予測確率とする．損失関数 l を用いた学習モデル $f : \mathcal{X} \to \mathcal{Y}$ の予測損失を $L^{std}(f) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim P_*}[l(\boldsymbol{y}, f(\boldsymbol{x}))]$ と定義し，訓練データ S を用いた経験損失を $L_n^{std}(f, S) := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim \widehat{P}_S}[l(\boldsymbol{y}, f(\boldsymbol{x}))]$ と定義する．ここで，$\mathbb{E}$ は添え字に示した確率分布に関する期待値を表す．

Mixup では，ニューラルネットワークの汎化性能を向上させるために，複数のサンプルを凸結合して混合することで新たなサンプルを作成する [1]．典型的には 2

[*1] 早稲田大学
Waseda University, Tokyo, Japan.

つのサンプルを混合する手法が用いられるため，本論文でもこれを前提とする．具体的には，サンプル $(\boldsymbol{x}, \boldsymbol{y})$, $(\boldsymbol{x}', \boldsymbol{y}')$ を混合して作成される新たなサンプル $(\widetilde{\boldsymbol{x}}, \widetilde{\boldsymbol{y}})$ は以下のように与えられる．

$$\widetilde{\boldsymbol{x}} := \lambda \boldsymbol{x} + (1-\lambda)\boldsymbol{x}', \quad \widetilde{\boldsymbol{y}} := \lambda \boldsymbol{y} + (1-\lambda)\boldsymbol{y}', \\ (\boldsymbol{x}, \boldsymbol{y}), (\boldsymbol{x}', \boldsymbol{y}') \sim \widehat{P}_S, \quad \lambda \sim Beta(\alpha, \alpha). \tag{1}$$

ここで，$\lambda \in [0,1]$ はベータ分布 $Beta(\alpha, \alpha)$ により定まる混合パラメータであり，$\alpha \in \mathbb{R}_+$ はベータ分布を規定するハイパーパラメータである．また，Mixup を用いた学習における経験損失は以下のように定義される：

$$L_n^{mix}(f, S) := \mathbb{E}_{mix}[l(\widetilde{\boldsymbol{y}}, f(\widetilde{\boldsymbol{x}}))], \tag{2}$$

ただし

$$\mathbb{E}_{mix}[\cdot] := \mathbb{E}_{(\boldsymbol{x},\boldsymbol{y})\sim\widehat{P}_S}[\mathbb{E}_{(\boldsymbol{x}',\boldsymbol{y}')\sim\widehat{P}_S}[\mathbb{E}_{\lambda\sim Beta(\alpha,\alpha)}[\cdot]]].$$

Mixup を提案した Zhang らは，L_n^{mix} を経験損失として最適化することは，訓練データ間における学習モデル f の振る舞いを線形化する効果をもたらすとしている [1]．

2.2 Mixup の解析

Mixup によって生成されたサンプルが元のサンプルの近傍にある場合に関しては，様々な研究が行われている [11, 5, 14]．これらの研究では，L_n^{mix} をテイラー展開することで，Mixup と正則化との関係について調べている．Carratino らは，損失関数を 2 階微分可能なクラスに限定し，L_n^{mix} を L_n^{std} と f の方向微分に依存する項に分解した [11]．Zhang らは，ある関数 h を用いて $l(\boldsymbol{y}, f(\boldsymbol{x})) = h(f(\boldsymbol{x})) - \boldsymbol{y}^\top f(\boldsymbol{x})$ のように損失関数が分解できるクラスに限定し，Mixup にヤコビアン正則化が含まれることを示した [5]．さらに，ロジスティック回帰などの学習アルゴリズムに対して，Mixup により adversarial robustness が改善されることを示した．Verma らは，クロスエントロピー損失とロジスティック損失に限定し，L_n^{mix} が J 階微分可能な f の J 次までの方向微分に依存する形で表せることを示した [14]．これらの解析では共通して，混合に使用する $\boldsymbol{x}$ または $\boldsymbol{x}'$ の周りでのテイラー展開が用いられている．このことは Mixup により作られたサンプル $\widetilde{\boldsymbol{x}}$ が $\boldsymbol{x}$ または $\boldsymbol{x}'$ の近傍にあることを暗黙的に仮定していることを意味する．さらにこの場合，$\widetilde{\boldsymbol{x}}$ は $\boldsymbol{x}$ または $\boldsymbol{x}'$ に微小な摂動を加えたものとみなすことができる．このことは，訓練データの入力にノイズを付与することとヤコビアン正則化が等価であること [6] にも関係していると考えられる．

訓練データの出力にあたるラベルに着目すると，Mixup によって生成されたサンプルが元のサンプルの近傍にある場合，$\widetilde{\boldsymbol{y}} = \boldsymbol{y} = \boldsymbol{y}'$ が成り立つか，$\widetilde{\boldsymbol{y}}$ が $\boldsymbol{y}$ または $\boldsymbol{y}'$ に微小な摂動を加えたものとみなせる．このことから，ラベルに一様なノイズを加える正則化手法である label smoothing [7] との関連性が議論されている．具体的には，[8, 9, 10, 11] では，Mixup と label smoothing の効果が，部分的に類似していることが示されている．例えば，Mixup と label smoothing が学習モデルの calibration を改善することや [8, 10]，予測確率のエントロピーを増大させること [11] が数値実験により確かめられている．また，Mixup が ECE を改善することが理論的に明らかにされている [9]．

一方，Mixup によって生成されたサンプルが元のサンプルの近傍にない場合に関する解析は，著者らの知る限りにおいてこれまでに行われていない．これは，Mixup に関する既存研究において用いられいるテイラー展開による解析が行えないことが原因である．本論文では，**3** 節において，$\widetilde{\boldsymbol{x}}$ が $\boldsymbol{x}$ または $\boldsymbol{x}'$ の近傍にない場合に，Mixup がヘッシアン正則化の効果を持つことを超平面で分離可能な二値分類データとロジスティック回帰において証明する．

3 主結果

本論文では，これまで解析されることのなかった，Mixup により生成された $\widetilde{\boldsymbol{x}}$ が $\boldsymbol{x}$ または $\boldsymbol{x}'$ の近傍以外にある場合について，超平面で分離可能な二値分類データとロジスティック回帰において解析し，ヘッシアン正則化の効果があることを証明する．**3.1** 節では解析する問題の設定を説明する．**3.2** 節では，設定した問題に基づいて Mixup にヘッシアン正則化の効果があることを示す．**3.3** 節では，より複雑な学習モデルにおいても同様の傾向があることを数値実験により確かめる．

3.1 問題設定

Mixup により生成された $\widetilde{\boldsymbol{x}}$ が $\boldsymbol{x}$ または $\boldsymbol{x}'$ の近傍以外にある場合，manifold intrusion [12] が生じる可能性がある．例えば，異なる 2 つのクラスのサンプルを凸結合する際，それらのサンプル間に別のクラスのサンプル群が存在すると manifold intrusion が生じやすくなる．この場合，Mixup が汎化性能の悪化を引き起こす可能性があるため [13] ，その効果を正確に調べることができない．そこで本論文では，manifold intrusion が生じない，超平面で分離可能な二値分類問題を考える．具体的には，分類すべきデータが従う分布として，次で定義される超平面で分離可能な 2 クラスデータを生成できる Gaussian model を設定する．

任意の $\boldsymbol{\theta}^* = (\theta_1^*, ..., \theta_D^*) \in \mathbb{R}_+^D, \sigma \in \mathbb{R}_+$ について，$(\boldsymbol{\theta}^*, \sigma)$-Gaussian model は，$(\boldsymbol{x}, y) \in \mathbb{R}^D \times \{-1, 1\}$ 上の確率分布として次のように定義される：

$$\boldsymbol{x}|y \sim \mathcal{N}(y \cdot \boldsymbol{\theta}^*, \sigma^2 \boldsymbol{I}), \quad \mathbb{P}(y) = \begin{cases} 1/2 & \text{if } y = 1 \\ 1/2 & \text{if } y = -1 \end{cases}. \tag{3}$$

Gaussian model は，mixup が ECE を改善することや，半教師あり学習が adversarial robustness を改善することなどの証明に用いられている [9, 15]．これらの研究 [9, 15] では，$\boldsymbol{\theta}^* \in \mathbb{R}^D$ としているが，本論文では簡単のため $\boldsymbol{\theta}^* \in \mathbb{R}_+^D$ としている．本論文の設定は，$\boldsymbol{\theta}^* \in \mathbb{R}^D$ の設定において適宜軸の符号を入れ替える変数変換を行った場合とみなせるため，解析する対象をより制限するものではない．以降では簡単のため，σ は既知とし，$\boldsymbol{\theta}^*$ のみが未知とする．また，2 つのクラスが超平面で分離可能である程度に σ は小さい値を持つとする．

この Gaussian model に独立に従う n 個の訓練データを $S = \{(\boldsymbol{x}_i, y_i)\}_{i=1,...,n}$ と定義する．このデータを学習する分類モデルとして，線形判別分析を用いる．線形判別分析による分類器 $C_{\boldsymbol{\theta}}$ は，訓練データ S を用いて，$C_{\boldsymbol{\theta}}(\boldsymbol{x}) := \mathrm{sign}(\boldsymbol{\theta}^\top \boldsymbol{x})$ と定義する．ただし，$\boldsymbol{\theta} = (\theta_1, \ldots, \theta_D) \in \mathbb{R}^D$ は分類器のパラメータである．訓練データ S を用いたときの $\boldsymbol{\theta}$ の最尤推定量 $\widehat{\boldsymbol{\theta}}$ は，$\widehat{\boldsymbol{\theta}} = \mathbb{E}_{(\boldsymbol{x},y)\sim\widehat{P}_S}[\boldsymbol{x}y]$ と求められる．また，訓練データ S に対して Mixup を適用したときの最尤推定量 $\widehat{\boldsymbol{\theta}}^{mix}(\alpha)$ は次

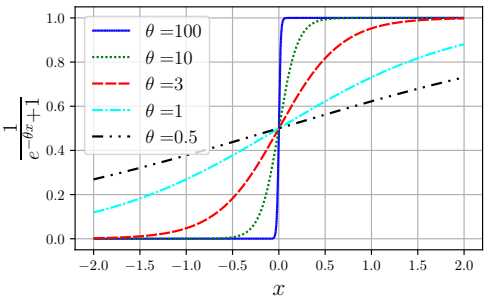

図 1　ロジスティック回帰モデル

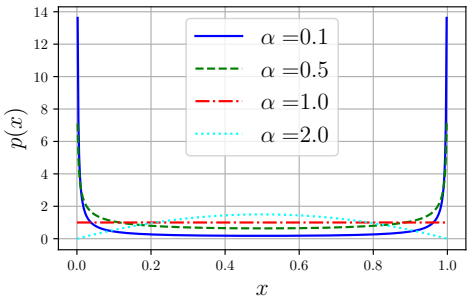

図 2　$Beta(\alpha, \alpha)$ の確率密度関数

のように計算される：

$$\widehat{\boldsymbol{\theta}}^{mix}(\alpha) = \mathbb{E}_{mix}[\widetilde{\boldsymbol{x}}\widetilde{y}]. \tag{4}$$

ただし，$\widetilde{\boldsymbol{x}} := \lambda\boldsymbol{x} + (1-\lambda)\boldsymbol{x}'$, $\widetilde{y} := \lambda y + (1-\lambda)y'$ である．

Mixup による学習モデルの振る舞いへの影響を調べるためには，各サンプルに対する学習モデルの予測確率を定義する必要がある．そこで，この分類器 $C_{\boldsymbol{\theta}}$ の確信度ベクトル $f_{\widehat{\boldsymbol{\theta}}}(\boldsymbol{x}) = (f_{\widehat{\boldsymbol{\theta}},1}(\boldsymbol{x}), f_{\widehat{\boldsymbol{\theta}},-1}(\boldsymbol{x}))^\top$ を，ロジスティック回帰モデルを用いて以下のように定義する：

$$f_{\boldsymbol{\theta},k}(\boldsymbol{x}) := \frac{1}{e^{-2k\cdot\boldsymbol{\theta}^\top\boldsymbol{x}/\sigma^2}+1}, \quad k \in \{1,-1\}. \tag{5}$$

確信度ベクトルの各要素を確信度スコアと呼ぶことにする．

3.2 Mixup とヘッシアン正則化の関係

3.2.1 モチベーション

$\widetilde{\boldsymbol{x}}$ が $\boldsymbol{x}$ または $\boldsymbol{x}'$ の近傍以外にある場合を調べるには，既存研究で用いられたテイラー展開のような訓練サンプルの近傍のみでの解析は使用できず，訓練データの凸包内全体での解析を行う必要がある．一方で，Mixup を提案した Zhang ら [1] によると，Mixup の主な効果は訓練サンプル間における学習モデルの振る舞いを線形にすることであると述べられている．本論文ではこの点に着目する．学習モデルの振る舞いが線形にならない可能性が高い部分は，主に異なるクラスのサンプル間であると考えられる．異なるクラスのサンプル間で予測確率が非線形に変化するモデル化が可能な例として，ロジスティック回帰モデルが挙げられる．実際，図 1 より，パラメータ θ の値がある程度大きい 1 変量ロジスティック回帰モデルは，予測確率が非線形に変化していることがわかる．また，図 1 からは，θ が小さいほどロジスティック回帰モデルの振る舞いが線形に近づいていることもわかる．したがって，ロジスティック回帰モデルで定義される確信度スコアに注目すれば，学習モデルの振る舞いがどれだけ線形に近いかを，パラメータ $\boldsymbol{\theta}$ を調べることで評価できるものと考えられる．

また，この評価を行うためには，学習モデルの振る舞いの線形性を測る指標が必要である．ある区間において学習モデルの振る舞いが線形であることは，その区間における学習モデルのヘッセ行列のすべての要素が 0 であることと等価である．よって，ある区間におけるヘッセ行列の大きさの和を指標として扱うことが考えられる．以上より，次のような解析の方針が得られる．仮に，パラメータ $\boldsymbol{\theta}$ が小さいほどロジスティック回帰モデルの振る舞いが線形に近づくならば，ロジスティック回帰モデルのヘッセ行列の大きさも小さくなると考えられる．これが成り立つならば，Mixup を適用した場合の推定量 $\widehat{\boldsymbol{\theta}}^{mix}(\alpha)$ と，Mixup を適用しない場合の推定量 $\widehat{\boldsymbol{\theta}}$ との大小関係を調べることで，Mixup にヘッシアン正則化の効果があるかを調べることができると考えられる．

3.2.2 理論解析

まず，**3.1** 節で定義した確信度スコア $f_{\boldsymbol{\theta},1}$ について，パラメータ $\boldsymbol{\theta}$ とヘッシアンの大きさの関係を調べる．そのために，評価すべき確信度スコア $f_{\boldsymbol{\theta},1}$ のヘッシアンの大きさを定義する．まず，確信度スコア $f_{\boldsymbol{\theta},1}$ の点 $\boldsymbol{x} \subseteq \mathcal{X}$ におけるヘッセ行列を $h(f_{\boldsymbol{\theta},1}(\boldsymbol{x}))$ と書く．次に，確信度スコアのヘッセ行列を評価する範囲 Θ^* を定める．Θ^* の最も単純な候補として，Mixup によりサンプルが作られる範囲を含む訓練データの凸包が考えられる．しかし，この凸包は訓練データに依存するため，同じ確信度スコアでも訓練データが異なると評価値が変わることから，評価する範囲には適さない．そこで，訓練データに依存せず，かつ確信度スコアが非線形になりやすい異なるクラスのサンプル間を十分に含むと考えられる範囲として，$\Theta^* := [-\theta_1^*, \theta_1^*] \times [-\theta_2^*, \theta_2^*] \times \cdots \times [-\theta_D^*, \theta_D^*]$ と定義する．この範囲に含まれない点は，決定境界から離れており，確信度スコアの変動が小さいと考えられることから，ヘッシアンの大きさの評価において重要ではないと考えられる．以上の定義を用いて，確信度スコア $f_{\boldsymbol{\theta},1}$ の評価すべきヘッシアンの大きさ $H(f_{\boldsymbol{\theta},1}, \boldsymbol{\theta}^*)$ を，次のように定義する：

$$H(f_{\boldsymbol{\theta},1}, \boldsymbol{\theta}^*) := \int_{\Theta^*} \|h(f_{\boldsymbol{\theta},1}(\boldsymbol{x}))\|_F^2 \mathrm{d}\boldsymbol{x}. \tag{6}$$

ただし，$\|\cdot\|_F$ はフロベニウスノルムとする．

この定義の下で，以下の定理が成り立つ．証明は，付録 A に記載する．

定理 1. 任意の次元 D を持つ $\mathcal{X}$ において $H(f_{\boldsymbol{\theta},1}, \boldsymbol{\theta}^*)$ は，$\boldsymbol{\theta}$ の各要素が正の範囲で，$\boldsymbol{\theta}$ の各要素に関して単調増加し，$\boldsymbol{\theta}$ の各要素が負の範囲で，$\boldsymbol{\theta}$ の各要素に関して単調減少する．

この定理より，Mixup を適用することで $H(f_{\boldsymbol{\theta},1}, \boldsymbol{\theta}^*)$ が小さくなるかどうかを調べるためには，Mixup の適用により $\boldsymbol{\theta}$ の推定量の各要素の絶対値が小さくなるかどうかを調べればよいことがわかる．以下の定理は，このことが成り立つことを保証するものである．証明は，付録 B に記載する．

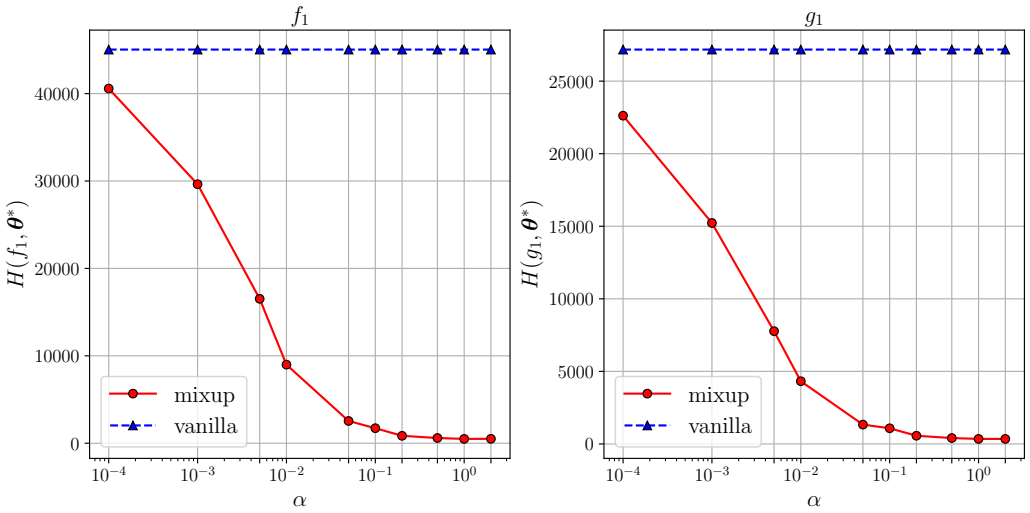

図 3　α と $H(f_1, \boldsymbol{\theta}^*), H(g_1, \boldsymbol{\theta}^*)$ の推移

定理 2. 任意の $\alpha \in \mathbb{R}_+$ に対して，

$$|\widehat{\theta}_i^{mix}(\alpha)| < |\widehat{\theta}_i|, \ \forall i \in \{1, \ldots, D\}.$$

が成り立つ．

定理 1，2 より，以下の系 1 が直ちに示される．系 1 は，Mixup がヘッシアン正則化の効果を持つことを示すものである．

系 1. 任意の $\alpha \in \mathbb{R}_+$ に対して，

$$H(f_{\widehat{\boldsymbol{\theta}}^{mix}(\alpha),1}, \boldsymbol{\theta}^*) < H(f_{\widehat{\boldsymbol{\theta}},1}, \boldsymbol{\theta}^*).$$

が成り立つ．

さらに，以下の系 2 が成り立つ．

系 2. 任意の $\alpha, \alpha' \in \mathbb{R}_+$ に対して，

$$\alpha < \alpha' \Rightarrow H(f_{\widehat{\boldsymbol{\theta}}^{mix}(\alpha'),1}, \boldsymbol{\theta}^*) \leq H(f_{\widehat{\boldsymbol{\theta}}^{mix}(\alpha),1}, \boldsymbol{\theta}^*).$$

が成り立つ．

系 2 は，α が大きいほど，ヘッシアン正則化の効果が大きくなることを示している．また，図 2 に示す Beta 分布の形状より，α が大きいほど，Mixup は，$\boldsymbol{x}, \boldsymbol{x}'$ の間の中央付近にサンプルを作りやすくなることがわかる．したがって，ヘッシアン正則化の効果は，Mixup により生成された $\widetilde{\boldsymbol{x}}$ が $\boldsymbol{x}$ または $\boldsymbol{x}'$ の近傍以外にある場合に得られるものであるといえる．

3.3 数値実験

本節では，Gaussian model から生成されたデータを，より複雑な学習モデルで学習した場合にも，**3.2** 節と同様の傾向があるかを数値実験により確かめる．学習モデルには次で定義される f, g を用いる．f は，幅が 20 の隠れ層を 4 層持ち，最終層としてシグモイド関数を持つニューラルネットワークとする．g は，幅が 20 の 2 層の全結合層と，1 層の全結合層を用いた残差接続の組み合わせを 3 つ繋げたニューラルネットワークとする．g の活性化関数には全て ReLU 関数を用いている．g の最終層は，幅が 20 の 1 層の全結合層とシグモイド関数であり，ResNet [16] を模した関数である．学習時のパラメータは，エポック数を 300，バッチサイズを 50，学習率を 0.01，モメンタムの係数を 0.9 とする．

データの生成には次のように設定した Gaussian model を用いる．式 (6) を計算する必要があるため，計算量の観点から，Gaussian model は 2 次元とする．Gaussian model のパラメータは，平面分離可能であるように，$\theta^* = [5, 5], \sigma = 1$ とし，各クラスのデータ数は 100 とする．式 (6) の計算には，積分区間において 500^D 点を調べ，数値積分により求めた．

図 3 に，α と，$H(f_1, \boldsymbol{\theta}^*), H(g_1, \boldsymbol{\theta}^*)$ との関係を示す．ただし，f_1, g_1 は f, g の $y = 1$ に関する確信度スコアとする．丸い点と実線が Mixup を用いた場合であり，三角の点と破線が Mixup を使用しない場合である．プロットした点は，3 回実験した結果の平均値を表している．図 3 より，複雑な学習モデルに Mixup を適用した場合もヘッシアン正則化の効果があることと，α を大きくするほどヘッシアン正則化の効果が大きいことがわかる．

4 結論

Mixup は，訓練サンプルを凸結合することにより新たなサンプルを作成するデータ拡張手法であり，多くの場面において汎化性能の向上に寄与してきた．本論文では，これまで解析されてこなかった Mixup で生成されるサンプルが元のサンプルから遠い場合についての理論解析を行った．その結果，manifold intrusion が生じない状況でのロジスティック回帰において，ヘッシアン正則化の効果があることを証明した．また，数値実験より，より複雑な DNN に Mixup を適用した場合にもヘッシアン正則化の効果が現れることを確認した．以上の結果を既存研究と組み合わせることで，ヤコビアン正則化，label smoothing，ヘッシアン正則化という複数の正則化手法をまとめて近似する手法として Mixup を特徴づけることができる．

謝辞

本研究の一部は，日本学術振興会における科学研究費補助金基盤研究（C）（課題番号 23K11111）による支援

を受けている．ここに記し謝意を表す．

参考文献

[1] Hongyi Zhang, Moustapha Cisse, Yann N Dauphin, and David Lopez-Paz. mixup: Beyond empirical risk minimization. In *International Conference on Learning Representations*, 2018.

[2] Linghui Meng, Jin Xu, Xu Tan, Jindong Wang, Tao Qin, and Bo Xu. Mixspeech: Data augmentation for low-resource automatic speech recognition. In *ICASSP 2021 - 2021 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7008–7012, 2021.

[3] Minghao Xu, Jian Zhang, Bingbing Ni, Teng Li, Chengjie Wang, Qi Tian, and Wenjun Zhang. Adversarial domain adaptation with domain mixup. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 6502–6509, 2020.

[4] Yiwei Wang, Wei Wang, Yuxuan Liang, Yujun Cai, and Bryan Hooi. Mixup for node and graph classification. In *Proceedings of the Web Conference 2021*, pages 3663–3674, 2021.

[5] Linjun Zhang and Zhun Deng. How does mixup help with robustness and generalization? In *The Ninth International Conference on Learning Representations*, 2021.

[6] Chris M Bishop. Training with noise is equivalent to tikhonov regularization. *Neural computation*, 7(1):108–116, 1995.

[7] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the inception architecture for computer vision. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2016.

[8] Sunil Thulasidasan, Gopinath Chennupati, Jeff A Bilmes, Tanmoy Bhattacharya, and Sarah Michalak. On mixup training: Improved calibration and predictive uncertainty for deep neural networks. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Linjun Zhang, Zhun Deng, Kenji Kawaguchi, and James Zou. When and how mixup improves calibration. In *International Conference on Machine Learning*, pages 26135–26160. PMLR, 2022.

[10] Rafael Müller, Simon Kornblith, and Geoffrey E Hinton. When does label smoothing help? *Advances in neural information processing systems*, 32, 2019.

[11] Luigi Carratino, Moustapha Cissé, Rodolphe Jenatton, and Jean-Philippe Vert. On mixup regularization. *arXiv preprint arXiv:2006.06049*, 2020.

[12] Hongyu Guo, Yongyi Mao, and Richong Zhang. Mixup as locally linear out-of-manifold regularization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 3714–3722, 2019.

[13] Zixuan Liu, Ziqiao Wang, Hongyu Guo, and Yongyi Mao. Over-training with mixup may hurt generalization. *arXiv preprint arXiv:2303.01475*, 2023.

[14] Vikas Verma, Sarthak Mittal, Wai Hoh Tang, Hieu Pham, Juho Kannala, Yoshua Bengio, Arno Solin, and Kenji Kawaguchi. Mixupe: Understanding and improving mixup from directional derivative perspective. *arXiv preprint arXiv:2212.13381*, 2022.

[15] Yair Carmon, Aditi Raghunathan, Ludwig Schmidt, John C Duchi, and Percy S Liang. Unlabeled data improves adversarial robustness. *Advances in neural information processing systems*, 32, 2019.

[16] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

付録 A 定理 1 の証明

まず，$D = 1$ の場合について証明する．

補題 1. $D = 1$ のとき，$H(f_{\theta,1}, \theta^*)$ は，$\theta > 0$ の範囲で θ に関して単調増加し，$\theta < 0$ の範囲で θ に関して単調減少する．

補題 *1* の証明. $v = 2/\sigma^2$ とおく．$D = 1$ のとき，

$$f'_{\theta,1}(x) = v\theta f_{\theta,1}(x)(1 - f_{\theta,1}(x)), \tag{7}$$

$$f''_{\theta,1}(x) = v^2\theta^2 f_{\theta,1}(x)(1 - f_{\theta,1}(x))(1 - 2f_{\theta,1}(x)) \tag{8}$$

であるので，$H(f_{\theta,1}, \theta^*)$ は次のようになる:

$$\begin{aligned} H(f_{\theta,1}, \theta^*) &= \int_{-\theta^*}^{\theta^*} (f''_{\theta,1}(x))^2 dx \\ &= \int_{-\theta^*}^{\theta^*} v^4\theta^4 (f_{\theta,1}(x))^2 (1 - f_{\theta,1}(x))^2 \\ &\qquad \cdot (1 - 2f_{\theta,1}(x))^2 \mathrm{d}x. \end{aligned} \tag{9}$$

$y = f_{\theta,1}(x)$ と変数変換すると，$\frac{\mathrm{d}x}{\mathrm{d}y} = \frac{1}{v}\left(\frac{1}{\theta y} + \frac{1}{\theta(1-y)}\right)$ であるので，$H(f_{\theta,1}, \theta^*)$ は，

$$\begin{aligned} H(f_{\theta,1}, \theta^*) &= \int_{f_{\theta,1}(-\theta^*)}^{f_{\theta,1}(\theta^*)} v^4\theta^4 y^2 (1-y)^2 (1-2y)^2 \\ &\qquad \cdot \frac{1}{v}\left(\frac{1}{\theta y} + \frac{1}{\theta(1-y)}\right) \mathrm{d}y \\ &= \begin{cases} v^3\theta^3(-\frac{1}{30} + f_{\theta,1}(\theta^*)^2 + 4f_{\theta,1}(\theta^*)^4) & \text{if } \theta > 0 \\ -v^3\theta^3(-\frac{1}{30} + f_{\theta,1}(\theta^*)^2 + 4f_{\theta,1}(\theta^*)^4) & \text{if } \theta < 0 \end{cases} \end{aligned} \tag{10}$$

と求められる．したがって，$f_{\theta,1}(\theta^*)$ は，$\theta > 0$ の範囲で θ に関して単調増加し，$\theta < 0$ の範囲で θ に関して単調減少することから，$H(f_{\theta,1}, \theta^*)$ は，$\theta > 0$ の範囲で θ に関して単調増加し，$\theta < 0$ の範囲で θ に関して単調減少する． □

定理 *1* の証明．$f_{\boldsymbol{\theta},1}(\boldsymbol{x})$ は，$\boldsymbol{\theta}^\top \boldsymbol{x}$ に関して単調増加する．よって，$f_{\boldsymbol{\theta},1}(\boldsymbol{x})$ は，$\boldsymbol{\theta}$ 方向にのみ増加する．そこで，$f_{\boldsymbol{\theta},1}(\boldsymbol{x})$ が 1 つの軸方向にのみ増加するような座標を求める．正規直交基底を $(\boldsymbol{e}_1,\dots,\boldsymbol{e}_D)$，求めるべき座標を $(\boldsymbol{e}'_1,\dots,\boldsymbol{e}'_D)$ と表す．$f_{\boldsymbol{\theta},1}(\boldsymbol{x})$ の増加方向を $\boldsymbol{e}'_1$ のみとするために，まず $\boldsymbol{e}'_1 := \sum_{i=1}^D \theta_i \boldsymbol{e}_i$ と定義する．$\boldsymbol{e}'_2,\dots,\boldsymbol{e}'_D$ は，

$$\boldsymbol{A} := \begin{bmatrix} \theta_1 & \xi_{12} & \cdots & \xi_{1D} \\ \vdots & \vdots & \ddots & \vdots \\ \theta_D & \xi_{D2} & \cdots & \xi_{DD} \end{bmatrix}, \quad \xi_{ij} \in \mathbb{R},\ \ i = 1,\dots,D, j = 2,\dots,D \tag{11}$$

を用いて，$[\boldsymbol{e}'_1,\dots,\boldsymbol{e}'_D] = [\boldsymbol{e}_1,\dots,\boldsymbol{e}_D]\boldsymbol{A}$ と定められるとする．$\boldsymbol{x}$ に対応する $(\boldsymbol{e}'_1,\dots,\boldsymbol{e}'_D)$ 上の点を $\boldsymbol{x}'$ とすると，

$$\begin{aligned} \boldsymbol{x} &= \sum_{i=1}^D x'_i \boldsymbol{e}'_i \\ &= x'_1 \sum_{j=1}^D \theta_j \boldsymbol{e}_j + \sum_{i=2}^D x'_i \Big(\sum_{j=1}^D \xi_{ji} \boldsymbol{e}_j \Big) \\ &= \sum_{j=1}^D \Big(\theta_j x'_1 + \sum_{i=2}^D \xi_{ji} x'_i \Big) \boldsymbol{e}_j \\ &= \boldsymbol{A}\boldsymbol{x}' \end{aligned} \tag{12}$$

と表すことができる．よって，

$$\begin{aligned} \boldsymbol{\theta}^\top \boldsymbol{x} &= \boldsymbol{\theta}^\top \boldsymbol{A}\boldsymbol{x}' \\ &= \Big(\sum_{i=1}^D \theta_i^2 \Big) x'_1 + \sum_{j=2}^D \Big(\sum_{i=1}^D \xi_{ij}\theta_i \Big) x'_j \end{aligned} \tag{13}$$

と変形できる．$f_{\boldsymbol{\theta},1}$ が x'_1 にのみ依存するような座標変換を行えばよいため，式 (13) より，$\sum_{i=1}^D \xi_{ij}\theta_i = 0,\ j = 2,\dots,D$ を満たし，かつ $\boldsymbol{e}'_2,\dots,\boldsymbol{e}'_D$ がそれぞれ異なるベクトルとなるように $\{\xi_{ij},\ i = 1,\dots,D, j = 2,\dots,D\}$ を定める．例えば，$D = 3$ のときは次のように $\boldsymbol{A}$ を定める：

$$\boldsymbol{A} := \begin{bmatrix} \theta_1 & -\frac{\theta_2+\theta_3}{\theta_1} & 1 & 1 \\ \theta_2 & 1 & -\frac{\theta_1+\theta_3}{\theta_2} & 1 \\ \theta_3 & 1 & 1 & -\frac{\theta_1+\theta_2}{\theta_3} \end{bmatrix}. \tag{14}$$

このような $\boldsymbol{A}$ の定め方は，任意の次元 D においても可能である．上記のように $\boldsymbol{A}$ を定めると，$\boldsymbol{e}'_1$ は $\boldsymbol{e}'_2,\dots,\boldsymbol{e}'_D$ と直交するため，変換後の座標系で考えれば，$f_{\boldsymbol{\theta},1}$ は x'_1 のみに依存する一変量関数とみなせる．このとき，x'_1 の係数は $\sum_{i=1}^D \theta_i^2$ である．また，$\boldsymbol{A}$ による座標変換後の積分範囲を，$\boldsymbol{e}'_1$ 方向の積分が一番外側になるように構成すると，正規直交基底上の $\boldsymbol{\theta}^*$ の $\boldsymbol{e}'_1$ 方向の成分を η と表せば，$\boldsymbol{e}'_1$ 方向の積分区間は $[-\eta,\eta]$ となる．$\boldsymbol{e}'_2,\dots,\boldsymbol{e}'_D$ 方向の積分は，$f_{\boldsymbol{\theta},1}$ は x'_1 のみに依存することから，各積分区間の大きさをかける操作となる．以上より，任意の次元 D における $H(f_{\boldsymbol{\theta},1})$ の計算は，$D = 1$ の場合の計算に帰着する．したがって，補題 1 より，定理 1 が示される．

□

付録 B 定理 2 の証明

定理 *2* の証明．$\mathbb{E}_{\lambda\sim Beta(\alpha,\alpha)}[\cdot]$ を $\mathbb{E}_\lambda[\cdot]$ と略記する．ベクトルに対する絶対値記号 $|\cdot|$ は，各要素について絶対値をとるものと定義する．また，ベクトルに対する不等号を，要素ごとに成り立つものと定義する．式 (4) より，$\widehat{\boldsymbol{\theta}}^{mix}(\alpha)$ について次が成り立つ：

$$\begin{aligned} &|\widehat{\boldsymbol{\theta}}^{mix}(\alpha)| \\ &= \left| \mathbb{E}_\lambda \left[\frac{1}{n^2} \sum_{i,j=1}^n (\lambda \boldsymbol{x}_i + (1-\lambda)\boldsymbol{x}_j)(\lambda y_i + (1-\lambda) y_j) \right] \right| \\ &= \left| \mathbb{E}_\lambda \left[(1 - 2\lambda(1-\lambda)) \frac{1}{n} \sum_{i=1}^n \boldsymbol{x}_i y_i + 2\lambda(1-\lambda)\boldsymbol{s} \right] \right| \\ &\le \frac{\alpha+1}{2\alpha+1} |\widehat{\boldsymbol{\theta}}| + \left(1 - \frac{\alpha+1}{2\alpha+1} \right) |\boldsymbol{s}|. \end{aligned} \tag{15}$$

ただし，$\boldsymbol{s} := (\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i)(\frac{1}{2n}\sum_{i=1}^n y_i)$ とする．最後の等式は，$\widehat{\boldsymbol{\theta}} = \frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i$ であることを用いた．データが超平面分離可能な程度に，Gaussian Model の分散パラメータ σ^2 が小さいと仮定していたことより，$|\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i| < |\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i y_i| = |\widehat{\boldsymbol{\theta}}|$ が成り立ち，また $|\frac{1}{2n}\sum_{i=1}^n y_i| < 1$ であるので，

$$|\boldsymbol{s}| = \left| \left(\frac{1}{n}\sum_{i=1}^n \boldsymbol{x}_i \right) \left(\frac{1}{2n}\sum_{i=1}^n y_i \right) \right| < |\widehat{\boldsymbol{\theta}}| \tag{16}$$

が成り立つ．よって，式 (15) より，$|\widehat{\boldsymbol{\theta}}^{mix}(\alpha)|$ は，$|\widehat{\boldsymbol{\theta}}|$ と $|\widehat{\boldsymbol{\theta}}|$ より各要素の値が小さい $|\boldsymbol{s}|$ との凸結合で表される．したがって，次の式が成り立つ：

$$|\widehat{\boldsymbol{\theta}}^{mix}(\alpha)| < |\widehat{\boldsymbol{\theta}}|. \tag{17}$$

□

付録 C 系 2 の証明

系 *2* の証明．式 (15) における $\frac{\alpha+1}{2\alpha+1}$ は，$\alpha \in \mathbb{R}_+$ に関して単調減少関数である．また，式 (15) と同様の計算により，$\widehat{\boldsymbol{\theta}}^{mix}(\alpha)$ について

$$\widehat{\boldsymbol{\theta}}^{mix}(\alpha) = \frac{\alpha+1}{2\alpha+1}\widehat{\boldsymbol{\theta}} + \left(1 - \frac{\alpha+1}{2\alpha+1} \right) \boldsymbol{s} \tag{18}$$

が成り立つ．よって，式 (16) と式 (18) より，α が大きいほど，$\widehat{\boldsymbol{\theta}}^{mix}(\alpha)$ は $\widehat{\boldsymbol{\theta}}$ より各要素の絶対値が小さい $\boldsymbol{s}$ に近づく．また，$\lim_{\alpha\to\infty} \frac{\alpha+1}{2\alpha+1} = \frac{1}{2}$ となることと式 (16) より，任意の $\alpha \in \mathbb{R}_+$ と $i = 1,\dots,D$ について，$\widehat{\theta}_i^{mix}(\alpha)$ はそれぞれ 0 と $\widehat{\theta}_i$ の間に値をとり得る．つまり，任意の $\alpha \in \mathbb{R}_+$ において $\widehat{\boldsymbol{\theta}}^{mix}(\alpha)$ と $\widehat{\boldsymbol{\theta}}$ の各要素の符号は等しい．よって，任意の $\alpha, \alpha' \in \mathbb{R}_+$ について，

$$\alpha < \alpha' \Rightarrow |\widehat{\boldsymbol{\theta}}_i^{mix}(\alpha')| < |\widehat{\boldsymbol{\theta}}_i^{mix}(\alpha)|, i = 1,\dots,D \tag{19}$$

が成り立つ．したがって，定理 1 より，任意の $\alpha, \alpha' \in \mathbb{R}_+$ について次が成り立つ：

$$\alpha < \alpha' \Rightarrow H(f_{\widehat{\boldsymbol{\theta}}^{mix}(\alpha'),1}, \boldsymbol{\theta}^*) \le H(f_{\widehat{\boldsymbol{\theta}}^{mix}(\alpha),1}, \boldsymbol{\theta}^*).$$

□