Hessian-based Analysis of Large Batch Training and 敵対者への堅牢性

Zhewei Yao1⇤ Amir Gholami1⇤ Qi Lei2 Kurt Keutzer1 Michael W. Mahoney1

1. カリフォルニア大学バークレー校、{zheweiy, amirgh, keutzer and mahoneymw}@berkeley.edu
2. テキサス大学オースティン校, leiqi@ices.utexas.edu 概要

ニューラルネットワークを大規模なバッチサイズで学習すると、現在の方法では精度が低下することがわかっています。この理由は、まだ完全には解明されていません。ここでは、ヘシアン演算子とロバスト最適化のレンズを通して、大規模なバッチサイズのトレーニングを研究します。

特に、ヘシアンに基づいた研究を行い、大規模なバッチサイズでトレーニングを行う際に損失関数のランドスケープがどのように変化するかを正確に分析します。本研究では、二階微分をバックプロパゲーションすることで、近似せずに真のヘシアン・スペクトルを計算します。複数のネットワークを対象とした広範な実験により、鞍部点は大規模なバッチサイズでの学習による汎化ギャップの原因ではなく、大規模なバッチは顕著に高いヘシアンスペクトルを持つ点に収束するという結果が一貫して示されている。さらに、ロバストな学習では、ヘシアン・スペクトルが大きい点は、敵対的な摂動に対するロバスト性が低いため、平坦な領域を優先することができることを示しています。さらに、この関係を研究し、ロバストな学習のための内部ループがサドルフリーの最適化問題であることを経験的・理論的に証明しました。本論文では、MNIST、 CIFAR-10、CIFAR-100 データセットを用いた、残差ネットワークを含む 5 つの異なるネットワークアーキテクチャの詳細な実験結果を示す。

1 はじめにニューラルネットワーク(NN)の学習では、入力データ **x** とそれに対応する

これは、超パラメータの選択が未知の分布 P から得られるサンプルに大きな影響を与える非凸最適化問題から得られたサンプルにラベルを付け、この未知の分布を学習するよう NN を訓練するものです。これは通常、P.実際には、一連の離散的な収束特性が観察されるだけである。特に、学習に大きなバッチサイズを用いると、収束特性の悪い点に収束してしまうことが多いことが観察されている。大規模なバッチを使用する主な動機

は，学習時間を短縮するために使用できるデータ並列化の機会が増えることです[13]．最近では，ラージバッチによる性能低下を回避するための様々な手法が提案されている[16, 28, 31]．しかし，これらの手法は，すべてのネットワークやデータセットで使えるわけではありません．このため，我々は元の問題を再検討し，大きなバッチサイズでの最適化が収束挙動にどのように影響するかを研究することにした．まず最初に、小バッチの学習中にヘシアンスペクトルと勾配がどのように変化するかを分析し、大バッチサイズと比較して、ロバストな学習との関連性を導き出します。特に、以下のような疑問に答えることを目的としています。

Q1 ラージバッチサイズでの学習は、スモールバッチサイズとどのように違うのでしょうか？同様に、ラージバッチサイズを使用した場合、モデルが収束する近傍の局所形状はスモールバッチに比べてどのような違いがありますか？

A1 2 次導関数をバックプロパゲートし、トレーニング中にそのスペクトルを計算します。この結果は，最適化を悩ませる鞍部点の存在に関する議論[6, 12]にもかかわらず，バッチ

サイズを勾配降下限界まで大きくしても，バッチサイズの大きい学習では実際には問題にならないことを示している．19]では，近似的な数値計算法を用いて

均等な貢献

32nd Conference on Neural Information Processing Systems (NeurIPS 2018), Montréal, Canada.

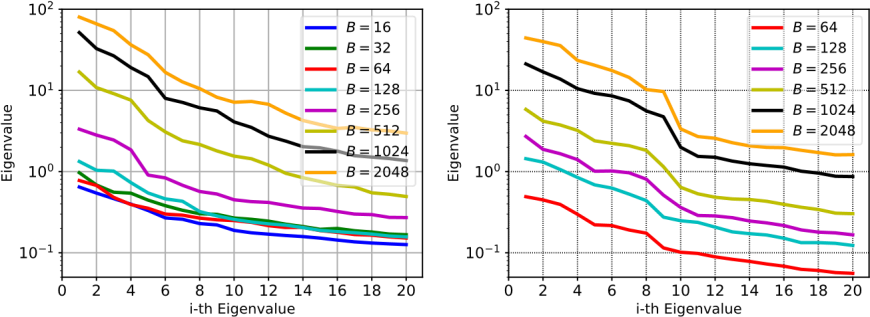


図 1: *CIFAR-10* の*C1(*左*)*と*MNIST* の*M1(*右*)*のデータセットにおけるヘシアンの上位*20* 個の固有値を示す．スペクトルは*power iteration* を用いて計算されており，相対誤差は1E-4 である．

が最大となる点があります。ここでは、真のヘシアンのスペクトルを直接計算することにより、バッチサイズが大きくなると、（支配的な固有値だけでなく）顕著に大きなスペクトルを持つ領域に徐々に閉じ込められていくことを示している。詳細は§2、特に図1、

2、4 を参照。

Q2 ロバストな最適化と大規模なバッチサイズのトレーニングとの関連性は？同様に、バッチサイズは、逆説的摂動に対するモデルのロバスト性にどのような影響を与えるのでしょうか？

A2 ロバスト最適化は、スペクトルの小さい領域（別名フラットミニマム）を好むという意味で、大規模なバッチトレーニングとは相反するものであることを示します。大きなバッチサイズで収束したポイントは、小さなバッチサイズで学習したモデルと比較して、敵対的な攻撃を受けやすくなることを示します。さらに、ロバストな学習を行うことで、逆にフラットなスペクトルを持つ点が得られ、敵対的な摂動に対してロバストになることを示します。最悪のケースを発見するロバスト最適化の内ループが、サドルフリー最適化問題であることを経験的・理論的に証明する。詳細は§3、特に表 1、7、図

4、6 で説明する。

限界私たちは、すべての論文には限界を明確に示すことが重要だと考えています。本研究では，最良の結果だけを報告することを避ける努力をし，すべての実験を少なくとも 3 回繰り返し，すべての発見が一貫していることを確認しました。さらに、複数のデータセットと残差ネットワークを含む複数のモデルでテストを行い、特定のテストに特化した結果を得ないようにしました。主な限界は、大規模なバッチトレーニングのためのソリューションを提案していないことです。ラージバッチとロバストなトレーニングの間に非常に有望な関係を示しているが、これは本来の問題を理解するための分析論文であることを強調している。これまでにいくつかの解決策が提案されていますが、それらは特定のケースでしか機能せず、大規模なハイパーパラメータのチューニングが必要です。

我々は、この論文の結果を大規模なバッチサイズのトレーニングの指針とするために、詳細なフォローアップ研究を行っています。

関連作品深層ニューラルネットワークは、幅広い用途で良好な性能を達成している。

DNN が使用できるさまざまな問題の多様性は，関数近似における効率に関連している[25,

7, 21, 1]．しかし、[32]の研究では、DNN が実際のデータセットで優れた性能を発揮するだけでなく、ランダムにラベル付けされたデータを非常によく記憶できることが示された。さらに，ネットワークの性能は，学習に用いるハイパーパラメータに大きく依存する．特に、最近の研究では、ニューラルネットワークは、入力データに知覚できないほどの摂動を加えると、簡単に騙されてしまうことがわかっています[15]。さらに，複数の研究により，大規模なバッチサイズの学習は，一般化能力が低いことがわかっています

[16, 31]．

ここでは、ニューラルネットワークの学習における後者の 2 つの側面に注目します。[19] は、大きなバッチは「よりシャープなミニマム」に収束するという結果を発表しました。

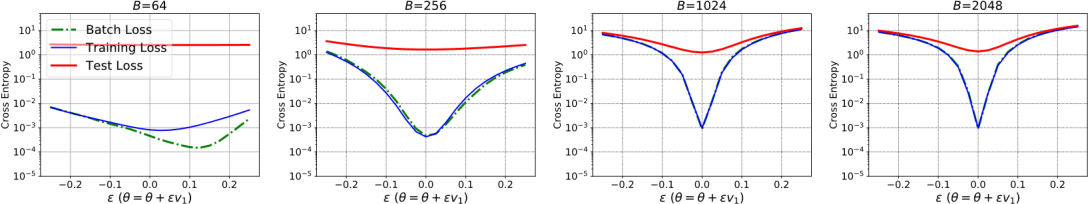
シャープなミニマムがフラットなミニマムと同じ学習損失を持つとしても，テストデータと学習データの間の小さな不一致は，容易に悪い汎化性能につながると主張された[19,

9]．フラットミニマム」がよく一般化するという事実は、[18]の初期の研究にさかのぼる。

著者らはフラットミニマムを最小記述長の理論[26]と関連づけ、フラットミニマムを実際に有利にする最適化手法を提案しました。最適化アルゴリズムを変更して「より良い」領域を見つけようとする同様の試みがいくつかあります [8, 5]。例えば、[5]はエントロピーSGD を提案しました。これは、ランジュバン・ダイナミクスを用いて損失関数を拡張し、

「エネルギー・ランドスケープ」の平坦な領域を好むようにしたものです。フラット/シャープネスの概念は、正確な定義がありません。異なるメトリクスの詳細な比較は[9]で議論されており、著者らはシャープな最小値もうまく一般化できることを示しています。

著者らはまた、重みの再パラメトリック化によってシャープネスを任意に変更できることを主張した。しかし、これは次のような場合には起こりません。

 図 2: *CIFAR-10* データセットにおける*C1* のヘシアンの支配的な固有ベクトルv1 に沿って、損失のランドスケープを示している。ここで✏は v1 に沿ってモデルパラメータを擾乱するスカラーである。

今回のケースでは，同じモデルを用いて，学習のハイパーパラメータを変更するだけである．28, 29]では，学習を確率微分方程式と見なすことを提案し，最適なバッチサイズは学習サイズと学習率に比例することを論じています．

我々の結果が示すように、NN がうまく機能しない場合を研究することで、相互に関連性があります。[30, 15]は、非常に優れた汎化能力を持つ NN を、入力にわずかな摂動を加えることで簡単に欺くことができることを発見しました。摂動の大きさは、ほとんどの場合、人間の目には感知できませんが、ネットワークの予測を完全に変えることができます。彼らは、Fast Gradient Sign Method (FGSM)と呼ばれる効果的な敵対的攻撃アルゴリズムを導入しました。そして、ニューラルネットワークの脆弱性を線形分類器と関連づけ、 RBF モデルは汎化性能がはるかに低いにもかかわらず、FGSM 攻撃に対してかなり頑健であることを示しました。その後、[20]では、FGSM 法が反復型 FGSM に拡張され、複数の勾配昇順ステップを実行して、敵対的摂動を計算するようになりました。反復 FGSM に基づく敵対的攻撃は，オリジナルの 1 ステップ FGSM よりも強力であることがわかった．敵対的攻撃に対抗するため、様々な防御策が提案されている[24, 14, 17, 2, 11]。モデルのロバスト性と大規模なバッチサイズの問題の間には、相互に関連性があることを後に示します。

本論文の構成は以下の通りである。まず，学習中にスペクトルがどのように変化するかを分析することで結果を示し，§2 では，異なるバッチサイズに対するモデルの一般化性能をテストする。セクション§3 では，どのようにして敵対的攻撃/トレーニングを行うかの詳細を議論する。特に、敵対的摂動を見つけることは、ある条件の下では saddle-free 問題であることを理論的に証明し、異なるバッチサイズに対するモデルのロバスト性をテストする。また、ロバストなトレーニングがスペクトラムにどのような影響を与えるかを実証的に示した結果を紹介する。最後に、セクション§4 では結論を述べる。

# 2 ラージバッチ、一般化ギャップ、ヘシアンスペクトル

セットアップ使用したネットワークのアーキテクチャを表 6 に示します。本文中では，この表で使われている略語で各アーキテクチャを表しています．特に指定のない限り，各バッチサイズは，学習損失が 0.001 以上になるまで学習される。異なるバッチは同じ条件で学習され，重みの減衰やドロップアウトは使用されていない．

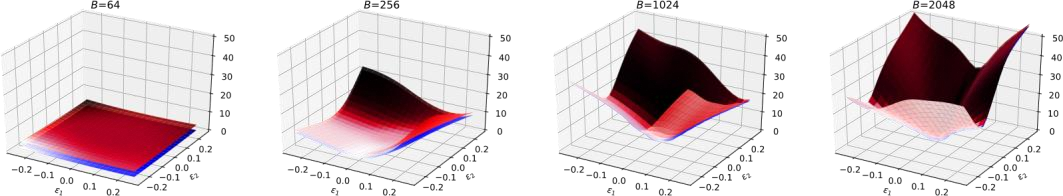
CIFAR-10 データセットの C1 ネットワークと MNIST の M1 ネットワークを対象に，大規模なバッチサイズで学習した結果をそれぞれ表 1，表 7 に示す．これを見ると，ある時点を

境に，バッチサイズを大きくするとテストデータセットの性能が低下することがわかる．これは，文献[19, 16]の結果と一致している．

前述したように、大規模なバッチサイズの一般化精度の低さに関する一般的な議論は、大規模なバッチが学習損失の「シャープな」最小値に引き付けられる傾向があるというものでした。19]では，与えられたモデルパラメータに対する損失関数の曲率を測定するために，近似的な指標が用いられた．ここでは，Hessian スペクトルを直接計算します．なお，Hessian 行列全体を計算することは，O(*N*2)行列であるため，実行不可能である．しかし、スペクトルは 23]の power iteration を用いて計算することができます。特に断りのない限り、個々の固有値の相対誤差が 1E-4 になるまで、Hessian の行列をパワーバックプ

ロパゲーションして反復を続けます。

この手法により、図 1 に示すように、異なるバッチサイズの Hessian の最初の上位 20 個の固有値を計算しました。さらに、支配的な固有値（✓1）の値を表 1 と表 2 に示します（LeNet-5 を用いた MNIST の結果は付録にあります。表 7 をご覧ください）。)図 1 から、すべての実験において、大きなバッチは、支配的な固有値と残りの 19 個の固有値の両方で、顕著に大きなヘシアンスペクトルを持っていることがわかります。しかし、曲率は非常に局所的な尺度であることに注意してください。それは、より

 図 3：*C1* モデルのパラメータを、摂動の大きさ✏1 と✏2 のヘシアンの最初の*2* つの支配的な固有ベクトルに沿って変化させたときの損失のランドスケープを示している。

損失関数が、モデルが収束した時点の周辺でどのように振る舞うかを調べることは有益です。このことを視覚的に示すために、C1 モデルと M1 モデルのそれぞれについて、モデルのパラメータを支配的な固有ベクトルに沿って変化させたときの総損失の変化を図 2 と図 7 に示しました。バッチサイズが大きいモデルは、テスト損失、トレーニング損失ともに、曲率の高い領域に引き寄せられていることがよくわかります。これは視覚的な図にも反映されています。また、3D プロットも追加しました。ここでは、図 3 に示すように、C1 モデルのパラメータを第 1 および第 2 の固有ベクトルの両方に沿って摂動させます。この視覚的な結果は、表 1 および表 7 のヘシアンスペクトル（✓1 参照）に示された数値と一致しています。例えば、表 1 の *B* = 512*,* 2048 のトレーニングおよびテスト損失の✓1 の値に注目し、図 3 の対応する結果を比較してください。

最近では、高次元のサドルポイントがニューラルネットワークの最適化を悩ませるという議論があります[6, 12]。我々は、学習中の全勾配とともに、ヘシアンの支配的な固有値を計算し、図 4 に報告しました。見ての通り、大きなバッチサイズは、より大きなスペクトルを持つ領域に徐々に引き寄せられていきますが、勾配がまだ大きいので、明らかにサドルポイントに行き詰まっていません。

表 1:*C1, C2* ネットワークを用いた*CIFAR-10* データセットの結果．異なるバッチトレーニングモデルのヘシアンスペクトルと、トレーニング*/*テストデータセットで生成された敵対的データセットでの対応する性能を示す（テスト結果は括弧内に記載）。

krxJk



BatchAcc.

✓

1

x1

Acc

✏

=0

*.*

02

Acc

✏

=0

*.*

01

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 16 | 100 (77.68) | 0.64 (32.78) | 2.69 (200.7) | 0.05 (20.41) | 48.07 (30.38) | 72.67 (42.70) |
| 32 | 100 (76.77) | 0.97 (45.28) | 3.43 (234.5) | 0.05 (23.55) | 49.04 (31.23) | 72.63 (43.30) |
| 64 | 100 (77.32) | 0.77 (48.06) | 3.14 (195.0) | 0.04 (21.47) | 50.40 (32.59) | 73.85 (44.76) |
| 128 | 100 (78.84) | 1.33 (137.5) | 1.41 (128.1) | 0.02 (13.98) | 33.15 (25.2 ) | 57.69 (39.09) |
| 256 | 100 (78.54) | 3.34 (338.3) | 1.51 (132.4) | 0.02 (14.08) | 25.33 (19.99) | 50.10 (34.94) |
| 512 | 100 (79.25) | 16.88 (885.6) | 1.97 (100.0) | 0.04 (10.42) | 14.17 (12.94) | 28.54 (25.08) |
| 1024 | 100 (78.50) | 51.67 (2372 ) | 3.11 (146.9) | 0.05 (13.33) | 8.80 (8.40 ) | 23.99 (21.57) |
| 2048 | 100 (77.31) | 80.18 (3769 ) | 5.18 (240.2) | 0.06 (18.08) | 4.14 (3.77 ) | 17.42 (16.31) |

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 256 | 100 (79.20) | 0.62 (28 ) | 12.10 (704.0) | 0.10 (41.95) | 0.57 (0.38) | 0.73 (0.47) |
| 512  1024    2048 | 100 (80.44) | 0.75 (57 ) | 4.82 (425.2) | 0.03 (26.14) | 0.34 (0.25) | 0.54 (0.38) |
| 100 (79.61) | 2.36 (142) | 0.523 (229.9) | 0.04 (17.16) | 0.27 (0.22) | 0.46 (0.35) |
| 100 (78.99) | 4.30 (307) | 0.145 (260.0) | 0.50 (17.94) | 0.18 (0.16) | 0.33 (0.28) |

# 3 ラージバッチ、アドバーサリアアタック、ロバストトレーニング

まず、敵対的攻撃とロバストな学習について簡単に説明し、次にこれらを大規模なバッチサイズの学習と結びつけた結果を示します。

3.1 ロバスト最適化とアドバーサリアアタック

ニューラルネットワークに対する敵対的攻撃の手法は、モデルのアーキテクチャやそのパラメータが既知であるホワイトボックス攻撃と、そのような情報が得られないブラックボックス攻撃に大別されます。ここでは、ホワイトボックス型の手法、特に最適化ベースのアプローチに焦点を当てて、攻撃と防御の両面から研究を進めています。

M(✓)を学習モデル（ニューラルネットワークアーキテクチャ）とし、(**x***,y*✓on)を入力データとし、(**x***,y*)を対応するラベルで表すとします。パラメータを持つネットワークの損失関数は

J が*(*✓*,***x***,y*)を最大化するようにする。逆襲のためには、摂動 J*(*✓*,***x***,y*)を求めます。 **x**

(with a bounded *L*1 or *L*2 norm) such max J*(*✓*,***x** + **x***,y),* (1) **x**2U

ここで、U は許容できる摂動の許容範囲です。このセットの典型的な選択は U 15] で、**x** を近似的に計算するための勾配法は Fast Gradient Sign Method [ the loss functional is computed w.r.t. input, and the perturbation is set to:

**x** (2) これは唯一の攻撃方法ではありません。他のアプローチとしては、反復的な FGSM 法

(*byFGSML*2*Grad*-10)[20]や、*L*1 ノルムの代わりに *L*2 ノルムなどの他のノルムを使用する（我々の結果では *L*2 法と表記している）。ここでは、入力に対するヘシアンを用いて、勾配の方向を 2 次の情報で事前に調整する 2 次攻撃も使用しています（詳細は付録の表 5 を参照してください）。このような攻撃に対しては，ロバストな学習を行うことが一つの方法である[30, 22]．

min max J*(*✓*,***x** + **x***,y).* (3)

✓**x**2U

この min-max 最適化問題を各反復で解くには、まず損失を最大化する最悪の敵対的摂動を見つけ、次にそれらのケースに対してモデルパラメータ✓を更新する必要があります。

敵対的な例は反復ごとに生成されなければならないため、目的関数を最大化する正確な摂動を見つけることは実行可能ではありません。その代わり、一般的な方法として、単一または複数の勾配上昇を実行して**x** を近似的に計算します。各反復で**x** を計算した後、典型的な最適化ステップ（SGD の変形）を実行して✓を更新します。

次に、最大化部分を解くことは、実際にはほとんどすべての場所でサドルフリー問題であることを示します。

この性質は、入力に対する Hessian が負の固有値を持たないことを意味し、これにより

§3.4 の 2 次の逆説的摂動テストのためのニュートン・ソルバーを実行するための CG です。 1

3.2 Adversarial perturbation:サドルフリー問題

我々の損失関数が J(✓; **x***,y*)であることを思い出してください。理論的な結果を示すために、モデルに対して以下のような仮定をします。

仮定 1.モデルの活性化関数は厳密に*ReLu* 活性化であり、すべての層は畳み込みまたは完全連結のいずれかであると仮定する。ここでは、バッチ・ノーマライゼーション層を受け入れます。なお、*ReLu* 活性化は原点で不連続、つまり*x*=0 であっても、*ReLu* 関数はほとんどの場所で*2* 回微分可能です。

次の定理は、以下を最大化する敵対的摂動を見つける問題を示しています。

は、どこにいても最大限の効果を発揮します。証明の詳細は付録を参照してください。 はサドルフリー最適化問題であり、入力 A.1 に対して PSD（Positive-Semi-Definite）Hessian を持ちます。

1 この結果は、GANS のより良い最適化戦略を見つけるのにも役立つかもしれません。

Theorem 1.仮定のもと*1* の場合、*DNN* の損失関数J*(*✓*,***x***,y*)は、ほぼすべての場所で、入力 **x** に対して鞍なし関数、すなわち

r2Jr*(*✓**x***,***2x***,y*) ⌫ 0*.*

定理 1 の証明から、DNN の次の命題がすぐに得られた。

命題 2.出力クラスが*c* 定理に基づいて、すなわち*y* 2{1*,*2*,*3*...,c1 with Assumption*}*A.1* であれば、*DNN* の*Hessian w.r.t. tofor* の詳細は以下の通り。*1*、入力**x**2*Rd* と**x** の数がほぼランク*c* の場合はの行列はほとんどどこにでも存在します。

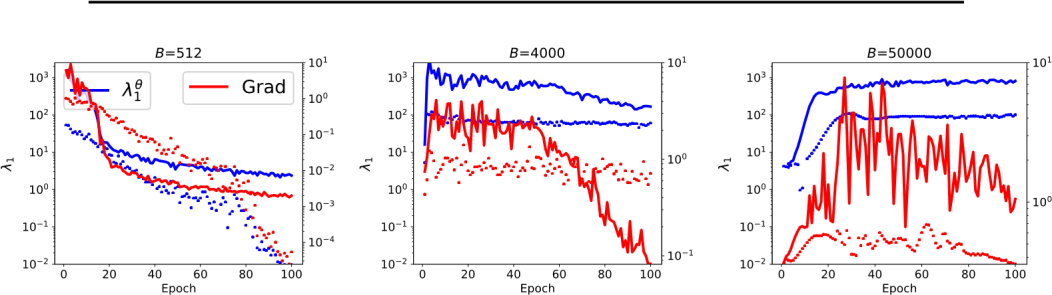
3.3 大規模バッチトレーニングとロバスト性

ここでは、異なるバッチで学習されたモデルの、敵対的な攻撃に対する堅牢性を検証します。すべての実験で高速グラディエント・サイン法を使用しています（以下の方法でも違いは見られませんでした）。

FGSM-10 攻撃）を行った。)敵対的な性能は，正しく分類された割合で測定される 表 2:*CR* ネットワークを用いた*CIFAR-100* データセットの結果．異なるバッチトレーニングモデルのヘシアンスペクトルと，トレーニング／テストデータセットで生成された敵対的データセットに対する対応する性能を示す（テスト結果は括弧内に記載）．

 BatchAcc. ✓1 Acc ✏ = 0*.*02 Acc✏= 0*.*01

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 64 | 99.98 (70.81) | 0.022 (10.43) | 61.54 (34.48) | 78.57 (39.94) |
| 128 | 99.97 (70.9 ) | 0.055 (26.50 ) | 58.15 (33.73) | 77.41 (38.77) |
| 256 | 99.98 (68.6 ) | 1.090 (148.29) | 39.96 (28.37) | 66.12 (35.02) |
| 512 | 99.98 (68.6 ) | 1.090 (148.29) | 40.48 (28.37) | 66.09 (35.02) |

 図 4：学習中のさまざまなエポックにおける，重みと全体の勾配に対するヘシアンの支配的な固有値の変化を示す．大規模なバッチと小規模なバッチでは、0（青い曲線）が増加していることに注意してください。特に、トータルグラジエントの値とヘシアンスペクトルは、ラージバッチがサドルポイントで「スタック」するのではなく、最適化ランドスケープの中で高い曲率を持つ領域でスタックすることを示しています。さらに詳しい結果を図 *16* に示します。点線は、ロバスト最適化を使用した場合の対応する結果を示しています。ロバスト最適化を使用すると、ソルバーはより小さなスペクトルの領域に留まるようになります。

逆説的な入力に対応しています。本研究では，様々な種類のデータを用いて，トレーニングデータセットとテストデータセットの両方について性能を評価した．

CIFAR-10 の C1 および C2 モデル、CIFAR-100 の CR モデルの性能結果は、✏=0*.*02*,*0*.*01 です（✏は*L*1 ノルムでの敵対的摂動の大きさを表す指標です）。 表 1,2 の最後の 2 列に報告されています（MNIST の結果は付録の表 7 に記載されています）。興味深いのは、すべてのケースにおいて、大規模なバッチは小規模なバッチに比べて敵対的攻撃を受けやすいということです。これは、モデルの設計だけでなく、最適化の際に使用されるハイパーパラメータや、特にモデルが収束したポイントの特性も、モデルのロバスト性に影響を与えることを意味します。

表 3：オリジナルモデル*MORI* に摂動を加えて得られた*MNIST* の様々な敵対的サンプルに対する各モデルの精度

*Dclean DFGSM DFGSM*

10

*DL*

2

*GRAD DFHSM DL*

2

*HESS*

アドバンストの平均値

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *MORI* | 99.32 | 60.37 | 77.27 | 14.32 | 82.04 | 33.21 | 53.44 |
| *MFGSM* | 99.49 | 96.18 | 97.44 | 63.46 | 97.56 | 83.33 | 87.59 |
| *MFGSM*10 | 99.5 | 96.52 | 97.63 | 66.15 | 97.66 | 84.64 | 88.52 |
| *ML*2*GRAD* | 98.91 | 96.88 | 97.39 | 86.23 | 97.66 | 92.56 | 94.14 |
| *MFHSM* | 99.45 | 94.41 | 96.48 | 52.67 | 96.89 | 77.58 | 83.60 |
| *L*2*HESS* | 98.72 | 95.02 | 96.49 | 77.18 | 97.43 | 90.33 | 91.29 |

M

この結果から、Hessian w.r.t*.* ✓のスペクトルとモデルのロバスト性には強い相関があるように思えます。しかし、一般的には Hessian w.r.t. weights と入力に対するモデルのロバスト性との間には相関がないことを強調したいと思います。例えば、2 変数の関数

J*(*✓*,***x***)(*✓と xw.r.t.を 2 つの単一変数として扱う)について、**x**.これは、最小二乗問題のヘシアンスペクトルを簡単に示すことができ、✓は *L* = k✓**x** yk22 のロバスト性とは無関係で

す。J ✓と **x** は、*xxT* .のヘシアンを見るのは難しくありません。

と✓*T* をそれぞれ設定しています。したがって、一般的には、重みを考慮したヘシアン・スペクトルをネットワークのロバスト性に結びつけることはできません。しかし、すべてのニューラルネットワークの数値結果は、Hessian spectrum w.r.t*.* ✓が高いモデルは、敵対的な攻撃を受けやすいことも示しています。これを説明するには、バッチサイズが異なる場合に、入力（すなわち **x**）に対する勾配とヘシアンがどのように変化するかを見ることが考えられます。我々は、学習データセットとテストデータセットの個々の入力サンプルについて、べき乗反復を用いて、このヘシアンの支配的な固有値を計算しました。

さらに、これらのデータセットについても、**x** に対する勾配のノルムを計算しました。これら 2 つの表 4：*CIFAR-10* の異なるサンプルに対する異なるモデルの精度（オリジナルモデル*MORI* に摂動を与えて得られたもの）。

*Dclean DFGSM DFGSM*

10

*DL*

2

*GRAD DFHSM DL*

2

*HESS*

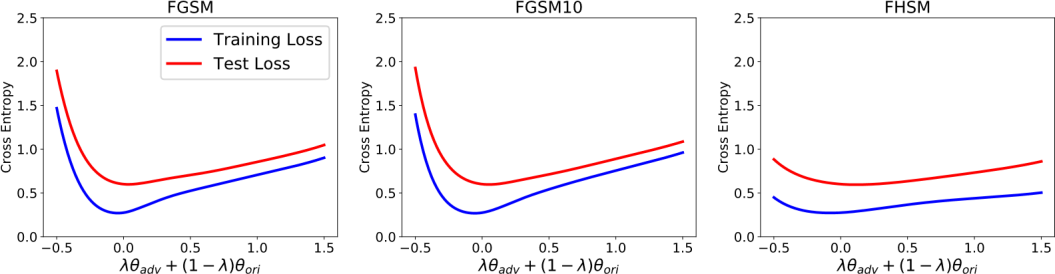
アドバンストの平均値

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| *MORI* | 79.46 | 15.25 | 4.46 | 12.37 | 29.64 | 22.93 | 16.93 |
| *MFGSM* | 71.82 | 63.05 | 63.44 | 57.68 | 66.04 | 62.36 | 62.51 |
| *MFGSM*10 | 71.14 | 63.32 | 63.88 | 58.25 | 65.95 | 62.70 | 62.82 |
| *ML*2*GRAD* | 63.52 | 59.33 | 59.73 | 57.35 | 60.44 | 58.98 | 59.16 |
| *MFHSM* | 74.34 | 47.65 | 43.95 | 38.45 | 62.75 | 55.77 | 49.71 |
| *L*2*HESS* | 71.59 | 50.05 | 46.66 | 42.95 | 62.87 | 58.42 | 52.19 |

M

のメトリクスが報告されていますが、これらの 2 つのメトリクスは実際には敵対的精度とは相関しないことがわかります。例えば，x1 と krxJk である．すべての実験の結果は、

*B* = 512 の C1 モデルを考慮しています。このモデルは、*B* = 32 と比較して、**x** に対する勾配とヘシアン固有値の両方が小さくなっていますが、敵対的な攻撃を受けた際の性能は、酸欠状態になっています。この理由として考えられるのは、大規模なバッチの決定境界が安定していないために、小さな敵対的摂動でモデルが騙されてしまうことです。

 図 5: *CIFAR-10* における*C3* モデルの*1-D* パラメトリックプロット。*MORI* と*MADV* のパラメータを補間し、*Y* 軸にクロスエントロピー損失を計算している。

3.4 Adversarial Training と Hessian Spectrum

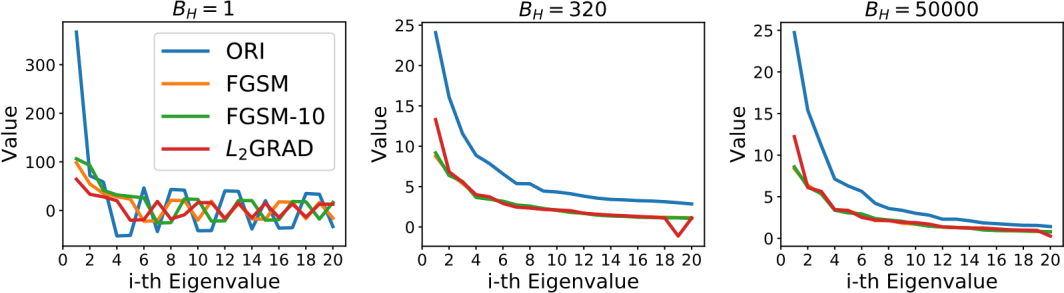
このパートでは、敵対的学習を行った後に、ヘシアンスペクトルと損失関数のランドスケープがどのように変化するかを研究する。ここでは、バッチサイズ（およびその他のすべての最適化ハイパーパラメータ）を固定し、§3.1 で説明した 5 つの異なる逆問題学習法を使用します。

わかりやすくするために、D をテストデータセットとする。テストデータセットには、オリジナルのクリーンなテストデータセット、敵対的なテストデータセット、または敵対的な手法を用いて作成されたテストデータセットがある。例えば、FGSM によって生成されたデータセットを *DFGSM* とし、オリジナルのクリーンなテストデータセットを *Dclean* とし

ます。

の設定を行いました．MNIST 実験では，MNIST データセット[3]を用いて標準的な LeNet を学習した（M1 ネットワークを使用）．最初の学習では，学習率を 0.01，運動量を 0.9 に設定し，5 エポック毎に学習率を半分に減少させ，合計 100 エポックを行った．さらに、学習率を 0*.*01 に設定して、5 エポック分の学習を行いました。摂動の大きさは

✏, は、*L*1 アタックでは 0*.*1、*L*2 アタックでは 2*.*8 に設定されています。また、C3 モデル[4] の結果も紹介します。

 図 6: 損失関数のサブサンプリングされたヘシアンのスペクトル（重みに応じて）．この結果は，ランダムに選ばれた異なるバッチサイズ，*B* = 1*,* 320*,* 50000 の *C1* に対して計算されたものである．

CIFAR-10 では，同じハイパーパラメータを用いて，100 回のエポックで学習を行っていることを除いて．その後、学習率 0*.*01、運動量 0*.*9（学習率は 5 エポック後に半分に減衰）で 10 エポックに渡って敵対的学習を行います。さらに、*L*1 アタックでは✏=0*.*02、*L*2 アタックでは 1*.*2 に設定しています[27]。

その結果を表 3、4 に示します。敵対的な学習を行うことで，これらの攻撃に対してモデルがより頑健になっていることがわかります．なお，敵対的攻撃の種類によって精度が異なるが，これは攻撃手法の強さが異なることから予想されることである．また、全ての敵対的学習手法は、*Dclean* では精度が低下するものの、敵対的データセットでは頑健性が向上しており、これは[15]の観察結果と一致している。例として、表 3 の 2 行目では、 FGSM をロバスト学習に使用した場合の結果を示しています。*L*2*GRAD* 攻撃手法に対するテストでは、オリジナルモデル（*MORI*）の14.32%に対し、このモデルの性能は 63.46%となりました。残りの行は、異なるアルゴリズムの結果を示しています。

ここでの主な疑問は、これらのロバスト最適化を行った後、損失関数のランドスケープがどのように変化するかということです。まず、図 5（すべてのケースで図 11 を参照）と図 10 に示すように、元のモデルのパラメータ✓とロバスト化されたモデルのパラメータの間の1 次元パラメトリック補間を示します。ロバスト化されたモデルは、オリジナルモデルに比べて曲率が小さい点にあることに注目してください。これを正確に定量化するために、図 6 と 12 に示すように、ヘシアンのスペクトルを計算します。完全なヘシアンのスペクトルの他に，サブサンプリングされたヘシアンのスペクトルも報告します．サブサンプリングされた Hessian のスペクトルは，訓練データセットのサブセットをランダムに選択して計算される．この部分集合のサイズを *BH* と呼ぶことで，学習バッチサイズとの混同を避けている．特に、*BH* = 1 と *BH* = 320 の結果を報告します。ここでいくつかの重要な発見があります。まず、ロバストモデルのスペクトルは、オリジナルモデルよりも明らかに小さいことに気付きます。これは、式 3 の min-max 問題が曲率の低い領域を好むことを意味しています。次に、トータル・ヘシアンは正の曲率を持つ点に収束したことを示していますが（少なくとも上位 20 個の固有値に基づく）、個々のサンプル（つまり *BH* = 1）を見ると、必ずしもそうではないことに注意してください。*BH* = 1 のバッチをランダムに選択すると、実際に正と負の曲率を持ち、勾配がゼロではない（つまりサドルポイントではない）点に収束していることがわかります。我々の知る限り、これは新しい発見ですが、SGD が個々の損失ではなく期待損失を最適化することから予想される結果です。

図 4 に戻って，ロバスト最適化を使用した場合のトレーニング中のスペクトルの変化を示します．ロバスト最適化を行うと、ロバストトレーニングを行わない場合と比較して、ソルバーがより小さなスペクトルの領域に押し込まれることがはっきりとわかります。これは非常に興味深い発見であり、シャープミニマムを避けるためにソルバーにバイアスをかける体系的な手段として、ロバスト・トレーニングを使用できる可能性を示しています。

# 4 結論

我々は、ヘシアン演算子のレンズを通してニューラルネットワークを研究しました。特に、大規模なバッチサイズの学習と、ホワイトボックス攻撃を受けた際のモデルの安定性との関係について研究しました。Hessian spectrum を計算することで、大規模なバッチサイズのトレーニングは Hessian spectrum の高い領域に引き寄せられる傾向があることを示すいくつかの証拠を提供しました。また，データセット全体に対するヘシアンの固有値を報告し，支配的な固有ベクトルに沿って擾乱を受けたときの損失の様子をプロットした．視覚的な結果は、スペクトルの数値と一致しました。実証実験の結果、敵対的攻撃・訓練とラージバッチが密接に関係していることがわかりました。私たちは、複数のデータセットを用いたいくつかの実証結果を提供し、大規模なバッチサイズのトレーニングが、より敵対的攻撃を受けやすいことを示しました（詳細な結果は補足資料に記載しています）。これは、モデル設計が重要であるだけでなく、最適化のハイパーパラメータがネットワークのロバスト性に大きく影響することを意味しています。さらに、ロバ

ストな学習は、大きなバッチサイズの学習とは相反するものであることがわかりました。

つまり、Hessian spectrum w.r.t*.* ✓が顕著に小さい領域が有利になるということです。その結果、モデルのロバスト性は、**x** に対するヘシアンとは（少なくとも直接的には）相関しないことがわかりました。また、このヘシアンは実際には PSD 行列であることがわかりました。つまり、敵対的摂動を見つける問題は、仮定 1 を満たす場合には実際にはサドルフリー問題であることがわかりました。さらに、全ての学習データを考慮した場合、モデルは正の曲率を持つ領域に収束しますが（つまり全損）、個々のサンプルを見た場合、ヘシアンは実際には大きな負の固有値を持つことがあることを示しました。これは、最適化の観点から、SGD がサンプルごとの損失ではなく期待損失を最適化していることに起因します。

# 5 反論

今回、レビューに時間を割いていただき、貴重なご意見をお寄せいただいたレビュアーおよびエリアチェアの皆様に感謝いたします。以下に、主なコメントを紹介します。

表 5：*C1* を用いた*SVHN* データセットの結果．*530K* 枚の画像を含む完全な学習データセットを使用している。バッチサイズが大きいという挙動は、他のデータセット（*CIFAR-10*、*CIFAR-100*、*MNIST*など）と一致している。

BatchAcc ✓1 Acc✏=0*.*05 Acc✏=0*.*02

256100

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 1024 | 100 (95.41 ) | 18.23 (184.7) | 16.65 (11.88) | 42.67 (29.27) |
| 4096 | 100 (95.22 ) | 58.46 (606.1) | 8.36 (6.33) | 27.04 (17.78) |

(95.70 ) 1.85 (87.71) 23.59

(16.46) 50.96 (35.89)

16384 100 (94.86 ) 74.28 (1040) 6.25 (4.95) 22.31 (15.43)

⇤敵対的正則化の予備的結果

(A) 本論文に続いて，我々は，敵対的学習（ロバストな最適化）と 2 次情報を組み合わせた新しい適応アルゴリズムを設計し，大規模なバッチ学習に対して最先端の性能を達成しました（[33]を参照）。この研究の主な目的は、大規模なバッチトレーニングの問題点

をよりよく理解するために詳細な分析を行うことでした。⇤ ReLU は 0 Hessian a.e.であり、

2 回微分可能な活性化による分析を加えることを提案します。

(A) これは、ReLU ネットワークに関する素晴らしい見解です。提案された活性化関数（Softplus と ELU）で新しい実験を行い、表 6 に結果を示しました。ReLU 活性化を選んだ理由は、新しいニューラルネットワークの多くがこれを取り入れているからです。しかし、私たちの結果は、2 回微分可能な活性化についても同様に当てはまります。つまり、より大きなバッチはロバスト性が低く（表 6 の最後の 2 列を参照）、最適化空間の中でより高いヘシアンスペクトル（シャープなポイントとしても知られる）を持つ領域に引き寄せられます。また、図 7 では、異なる活性化関数のバッチサイズに対するヘシアンの支配的な固有値を視覚的にプロットしています。これも同じ傾向を示しています。これらの結果は、論文の最終版に追加する予定です。⇤ 2 つの小さなデータセット MNIST と

CIFAR-10 での実験。

(A) 我々の結果は、この 2 つの小さなデータセットに限定されるものではありません。表 5 に示すように，53 万枚の画像を含む SVHN データセットを用いて実験を行い，レビューアの懸念に応えました．結果は、他のデータセットと一致しており、バッチが大きいほど敵対的摂動に対するロバスト性が低く（最後の 2 列）、ヘシアン・スペクトルもバッチが大きいほど大きくなることがわかります。

*(*✓1 欄をご覧ください）。)

BatchAcc.

✓1  2048100.00 (76.27 ) 97.71 (4329)

128100.00 (78.79 ) 4.45 (318.9) 128100.00 (78.94 ) 4.32 (271.4)

256100.00 (78.79 ) 5.00 256100.00 (78.88 ) 17.39 (469.2)

(507.4) 512100.00 (78.38 ) 27.23 (1048) 1024100.00 (77.82 ) 62.64

512100.00 (78.68 ) 16.18 (819.2) (2392)

1024100.00 (77.78 ) 46.99 2048100.00 (76.64 ) 114.4 (4347)

(2030)

表 6:*CIFAR-10* データセットにおける *C1S*

*(ReLU* を *Softplus* に置き換える, =20*)* と

*C1E (ReLU* を *ELU* に置き換える, ↵ =1*)* によ 2.38 (2.29)

る結果。これらの*2*回微分可能なアクテ 17.37 (15.26)ィベーションでも、*ReLU* の場合と同じ 13.44 (12.01) 傾向が見られる。 9.20 (8.74)

Acc✏=0*.*02 4.10 (3.99) 2 1.55 (1.6)

0

.

図 7：様々なバッチサイズで異なる活性化

( 関数を用いた*C1* の訓練データセットにおけ

1

る上位固有値。

# 7 リファレンス

1. マーティン・アンソニー、ピーター・L・バートレット *Neural Network*

*Learning:Theoretical Foundations*. cambridge university press, 2009.

1. Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal.機械学習分類器への回避攻撃に対する防御としての次元削減. *arXiv preprint arXiv:1704.02654*, 2017.
2. Léon Bottou, Corinna Cortes, John S Denker, Harris Drucker, Isabelle Guyon, Lawrence D Jackel, Yann LeCun, Urs A Muller, Edward Sackinger, Patrice Simard, et al. 分類法の比較：手書き数字認識のケーススタディ.コンピュータビジョン＆画像処理、第*12* 回*IAPR* 国際会議の議事録、第 2 巻、77-82 ページ、1994 年。
3. ニコラス・カルリーニ、デビッド・ワグナーTowards evaluating the robustness of neural networks.*Security and Privacy (SP*), pages 39-57.IEEE, 2017.



.79) 17.79 (16.19)

12.99 (11.55)

5.55 (5.42)

1. Pratik Chaudhari, Anna Choromanska, Stefano Soatto, and Yann LeCun.Entropy-SGD: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.
2. Yann N Dauphin, Razvan Pascanu, Caglar Gulcehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio.高次元非凸最適化におけるサドルポイント問題の識別と攻撃.In *Advances in neural information processing systems*, pages 2933-2941, 2014.
3. オリビエ・デラレオとヨッシャ・ベンギオ。Shallow vs. Deep Sum-product networks.In

*Advances in Neural Information Processing Systems*, pages 666-674, 2011.

1. Guillaume Desjardins, Karen Simonyan, Razvan Pascanu, and Koray Kavukcuoglu.ナチュラルニューラルネットワーク。In *Advances in Neural Information Processing Systems*, pages 2071-2079, 2015.
2. Laurent Dinh, Razvan Pascanu, Samy Bengio, and Yoshua Bengio.Sharp minima can generalize for deep nets. *arXiv preprint arXiv:1703.04933*, 2017.
3. Stanley C Eisenstat and Homer F Walker.非正確なニュートン法における強制項の選択。

*SIAM Journal on Scientific Computing*, 17(1):16-32, 1996.

1. Reuben Feinman, Ryan R Curtin, Saurabh Shintre, and Andrew B Gardner.Detecting adversarial samples from artifacts. *arXiv preprint arXiv:1703.00410*, 2017.
2. Rong Ge, Furong Huang, Chi Jin, and Yang Yuan.Escaping from saddle points-online stochastic gradient for tensor decomposition.In *Conference on Learning Theory*, pages 797- 842, 2015.
3. A.Gholami, A. Azad, P. Jin, K. Keutzer, and A. Buluc.ニューラルネットワークの学習におけ

るモデル、バッチ、ドメインの並列性の統合。*SPAA'8*, 2018」を発表しました。[PDF】をご覧ください。]

1. Zhitao Gong, Wenlu Wang, and Wei-Shinn Ku.Adversarial and clean data are not twins. *arXiv preprint arXiv:1704.04960*, 2017.
2. Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy.敵対的な例を説明して利用する。 *arXiv preprint arXiv:1412.6572*, 2014.
3. Priya Goyal, Piotr Dollár, Ross Girshick, Pieter Noordhuis, Lukasz Wesolowski, Aapo Kyrola, Andrew Tulloch, Yangqing Jia, and Kaiming He.Accurate, large minibatch SGD: training imagenet in 1 hour. *arXiv preprint arXiv:1706.02677*, 2017.
4. Kathrin Grosse, Praveen Manoharan, Nicolas Papernot, Michael Backes, and Patrick McDaniel.On the (statistical) detection of adversarial examples. *arXiv preprint arXiv:1702.06280*, 2017.
5. Sepp Hochreiter と Jürgen Schmidhuber。フラットミニマム。*Neural Computation*, 9(1):1-

42, 1997.

1. Forrest N Iandola, Matthew W Moskewicz, Khalid Ashraf, and Kurt Keutzer.Firecaffe: nearlinear acceleration of deep neural network training on compute clusters.In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2592-2600, 2016.
2. Nitish Shirish Keskar, Dheevatsa Mudigere, Jorge Nocedal, Mikhail Smelyanskiy, and Ping Tak Peter Tang.On large-batch training for deep learning:Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836*, 2016.
3. Alexey Kurakin, Ian Goodfellow, and Samy Bengio.物理的世界における敵対的事例. *arXiv preprint arXiv:1607.02533*, 2016.
4. Nicolas Le Roux と Yoshua Bengio。深層信念ネットワークはコンパクトな普遍的近似器である。*Neural computation*, 22(8):2192-2207, 2010.
5. Jason D Lee, Ioannis Panageas, Georgios Piliouras, Max Simchowitz, Michael I Jordan, and Benjamin Recht.一次の手法はほとんど常にサドルポイントを回避する。*arXiv preprint arXiv:1710.07406*, 2017.
6. Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu.Towards deep learning models resistant to adversarial attacks.学習表現に関する国際会議、2018。
7. James Martens, Ilya Sutskever.ヘシアンフリー最適化による深層およびリカレントネットワークのトレーニング。In *Neural Networks:*トリックオブザトレード、ページ 479-

535。Springer, 2012.

1. Jan Hendrik Metzen, Tim Genewein, Volker Fischer, and Bastian Bischoff.On detecting adversarial perturbations. *arXiv preprint arXiv:1702.04267*, 2017.
2. Guido F Montufar, Razvan Pascanu, Kyunghyun Cho, and Yoshua Bengio.On the number of linear regions of deep neural networks.In *Advances in neural information processing systems*, pages 2924-2932, 2014.
3. Jorma Rissanen.最短のデータ記述によるモデリング.*Automatica*, 14(5):465-471, 1978.
4. Uri Shaham, Yutaro Yamada, and Sahand Negahban.敵対的なトレーニングを理解する。

Increasing local stability of neural nets through robust optimization. *arXiv preprint arXiv:1511.05432*, 2015.

1. Samuel L Smith, Pieter-Jan Kindermans, and Quoc V Le.Don't decay the learning rate, increase the batch size. *arXiv preprint arXiv:1711.00489*, 2017.
2. Samuel L Smith and Quoc V Le.A bayesian perspective on generalization and stochastic gradient descent.*Second workshop on Bayesian Deep Learning (NIPS 2017*), 2017.
3. Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus.ニューラルネットワークの魅力的な特性. *arXiv preprint arXiv:1312.6199*, 2013.
4. Zhewei Yao, Amir Gholami, Kurt Keutzer, and Michael Mahoney.Large batch size training of neural networks with adversarial training and second-order information. *arXiv preprint arXiv:1810.01021*, 2018.
5. Yang You, Igor Gitman, and Boris Ginsburg.Scaling SGD batch size to 32K for imagenet training. *arXiv preprint arXiv:1708.03888*, 2017.
6. Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals.Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530*, 2016.