

# On Large-Batch Training for Deep Learning: Generalization Gap and Sharp Minima

Nitish Shirish Keskar<sup>1</sup> Dheevatsa Mudigere<sup>2</sup> Jorge Nocedal<sup>1</sup>  
Mikhail Smelyanskiy<sup>2</sup> Ping Tak Peter Tang<sup>2</sup>

<sup>1</sup>Northwestern University

<sup>2</sup>Intel Corporation

ICLR, 2017

Presenter: Tianlu Wang

- 1 Introduction
  - Batch Size of Stochastic Gradient Methods
- 2 Drawbacks of Large-Batch Methods
  - Main Observation
  - Numerical Results
  - Parametric Plots
  - Sharpness of Minima
- 3 Success of Small-Batch Methods
  - Deterioration along Increasing of Batch-Size
  - Warm-started Large Batch experiments
- 4 Summary

# Outline

## 1 Introduction

- Batch Size of Stochastic Gradient Methods

## 2 Drawbacks of Large-Batch Methods

- Main Observation
- Numerical Results
- Parametric Plots
- Sharpness of Minima

## 3 Success of Small-Batch Methods

- Deterioration along Increasing of Batch-Size
- Warm-started Large Batch experiments

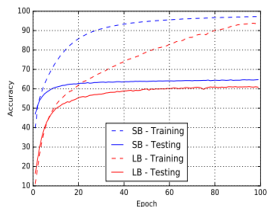
## 4 Summary

# Batch Size of Stochastic Gradient Methods

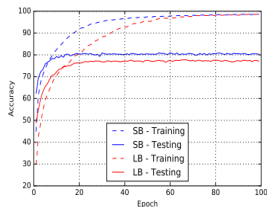
- Non-convex optimization in deep learning:  
$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{M} \sum_{i=1}^M f_i(x)$$
- Stochastic Gradient Methods and its variants:  
 $|B_k| \in \{32, 64, \dots, 512\}$
- Increase batch size to improve parallelism leads to **a loss in generalization performance**

# Batch Size of Stochastic Gradient Methods

- Non-convex optimization in deep learning:  
$$\min_{x \in \mathbb{R}^n} f(x) := \frac{1}{M} \sum_{i=1}^M f_i(x)$$
- Stochastic Gradient Methods and its variants:  
 $|B_k| \in \{32, 64, \dots, 512\}$
- Increase batch size to improve parallelism leads to **a loss in generalization performance**



(a) Network  $F_2$



(b) Network  $C_1$

Figure 2: Training and testing accuracy for SB and LB methods as a function of epochs.

# Outline

- 1 Introduction
  - Batch Size of Stochastic Gradient Methods
- 2 Drawbacks of Large-Batch Methods
  - **Main Observation**
  - Numerical Results
  - Parametric Plots
  - Sharpness of Minima
- 3 Success of Small-Batch Methods
  - Deterioration along Increasing of Batch-Size
  - Warm-started Large Batch experiments
- 4 Summary

# Main Observations

- Large-batch methods tend to converge to **sharp minimizers** of the training function and tend to generalize less well. Small-batch methods converge to **flat minimizers** and are able to escape basins of attraction of sharp minimizers.

# Main Observations

- Large-batch methods tend to converge to **sharp minimizers** of the training function and tend to generalize less well. Small-batch methods converge to **flat minimizers** and are able to escape basins of attraction of sharp minimizers.
- Sharp Minimizer  $\hat{x}$ : function increases rapidly in a small neighborhood of  $\hat{x}$   
Flat Minimizer  $\bar{x}$ : function varies slowly in a large neighborhood of  $\bar{x}$

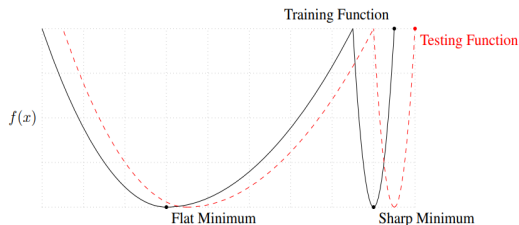


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)



# Outline

- 1 Introduction
  - Batch Size of Stochastic Gradient Methods
- 2 Drawbacks of Large-Batch Methods
  - Main Observation
  - **Numerical Results**
  - Parametric Plots
  - Sharpness of Minima
- 3 Success of Small-Batch Methods
  - Deterioration along Increasing of Batch-Size
  - Warm-started Large Batch experiments
- 4 Summary

# Numerical Results

- 6 multi-class classification networks, mean cross entropy, ADAM optimizer, LB: 10% of training data, SB: 256 data points

# Numerical Results

- 6 multi-class classification networks, mean cross entropy, ADAM optimizer, LB: 10% of training data, SB: 256 data points

Table 1: Network Configurations

Name	Network Type	Architecture	Data set
$F_1$	Fully Connected	Section B.1	MNIST (LeCun et al., 1998a)
$F_2$	Fully Connected	Section B.2	TIMIT (Garofolo et al., 1993)
$C_1$	(Shallow) Convolutional	Section B.3	CIFAR-10 (Krizhevsky & Hinton, 2009)
$C_2$	(Deep) Convolutional	Section B.4	CIFAR-10
$C_3$	(Shallow) Convolutional	Section B.3	CIFAR-100 (Krizhevsky & Hinton, 2009)
$C_4$	(Deep) Convolutional	Section B.4	CIFAR-100

# Numerical Results

- 6 multi-class classification networks, mean cross entropy, ADAM optimizer, LB: 10% of training data, SB: 256 data points

Table 1: Network Configurations

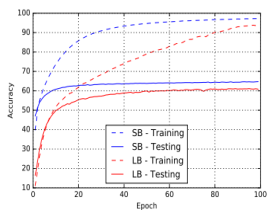
Name	Network Type	Architecture	Data set
$F_1$	Fully Connected	Section B.1	MNIST (LeCun et al., 1998a)
$F_2$	Fully Connected	Section B.2	TIMIT (Garofolo et al., 1993)
$C_1$	(Shallow) Convolutional	Section B.3	CIFAR-10 (Krizhevsky & Hinton, 2009)
$C_2$	(Deep) Convolutional	Section B.4	CIFAR-10
$C_3$	(Shallow) Convolutional	Section B.3	CIFAR-100 (Krizhevsky & Hinton, 2009)
$C_4$	(Deep) Convolutional	Section B.4	CIFAR-100

Name	Training Accuracy		Testing Accuracy	
	SB	LB	SB	LB
$F_1$	99.66% $\pm$ 0.05%	99.92% $\pm$ 0.01%	98.03% $\pm$ 0.07%	97.81% $\pm$ 0.07%
$F_2$	99.99% $\pm$ 0.03%	98.35% $\pm$ 2.08%	64.02% $\pm$ 0.2%	59.45% $\pm$ 1.05%
$C_1$	99.89% $\pm$ 0.02%	99.66% $\pm$ 0.2%	80.04% $\pm$ 0.12%	77.26% $\pm$ 0.42%
$C_2$	99.99% $\pm$ 0.04%	99.99% $\pm$ 0.01%	89.24% $\pm$ 0.12%	87.26% $\pm$ 0.07%
$C_3$	99.56% $\pm$ 0.44%	99.88% $\pm$ 0.30%	49.58% $\pm$ 0.39%	46.45% $\pm$ 0.43%
$C_4$	99.10% $\pm$ 1.23%	99.57% $\pm$ 1.84%	63.08% $\pm$ 0.5%	57.81% $\pm$ 0.17%

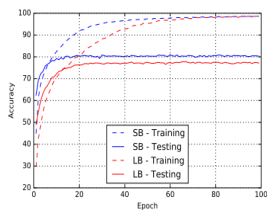
# Question

- Generalization gap is not due to *over-fitting* or *over-training* ???

- Generalization gap is not due to *over-fitting* or *over-training* ???



(a) Network  $F_2$



(b) Network  $C_1$

Figure 2: Training and testing accuracy for SB and LB methods as a function of epochs.

# Outline

- 1 Introduction
  - Batch Size of Stochastic Gradient Methods
- 2 Drawbacks of Large-Batch Methods
  - Main Observation
  - Numerical Results
  - **Parametric Plots**
  - Sharpness of Minima
- 3 Success of Small-Batch Methods
  - Deterioration along Increasing of Batch-Size
  - Warm-started Large Batch experiments
- 4 Summary

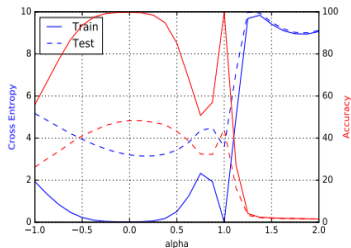
# Parametric Plots

- $x_s^*$  and  $x_j^*$ : solutions obtained by SB and LB
- plot  $f(\alpha x_j^* + (1 - \alpha)x_s^*)$ :

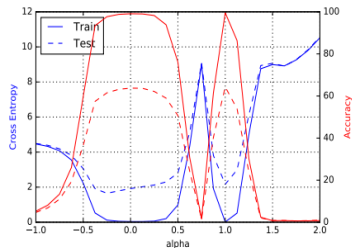


# Parametric Plots

- $x_S^*$  and  $x_I^*$ : solutions obtained by SB and LB
- plot  $f(\alpha x_I^* + (1 - \alpha)x_S^*)$ :



(e)  $C_3$



(f)  $C_4$

# Outline

- 1 Introduction
  - Batch Size of Stochastic Gradient Methods
- 2 Drawbacks of Large-Batch Methods
  - Main Observation
  - Numerical Results
  - Parametric Plots
  - Sharpness of Minima
- 3 Success of Small-Batch Methods
  - Deterioration along Increasing of Batch-Size
  - Warm-started Large Batch experiments
- 4 Summary

# Sharpness of Minima

- Motivation: Measure the sensitivity of training function at the given local minimizer, so we want to explore **a small neighborhood** of a minimizer and compute **the largest value** that  $f$  can attain in this neighborhood.

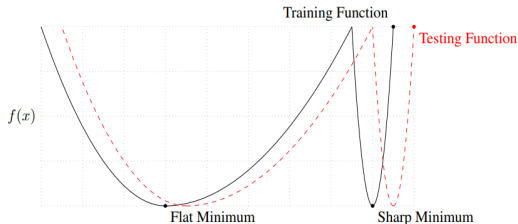


Figure 1: A Conceptual Sketch of Flat and Sharp Minima. The Y-axis indicates value of the loss function and the X-axis the variables (parameters)

# Sharpness of Minima

- Small neighborhood:

$p$ : dimension of manifold

$A$ :  $n \times p$  matrix, columns are randomly generated

$A^+$ : pseudo-inverse of  $A$

$$C_\varepsilon = \{z \in \mathbb{R}^n : -\varepsilon(|x_i| + 1) \leq z_i \leq \varepsilon(|x_i| + 1)\} \\ \forall i \in \{1, 2, \dots, n\}$$

$$C_\varepsilon = \{z \in \mathbb{R}^p : -\varepsilon(|(A^+x)_i| + 1) \leq z_i \leq \varepsilon(|(A^+x)_i| + 1)\} \\ \forall i \in \{1, 2, \dots, p\}$$

- **Metric 2.1.** Given  $x \in \mathbb{R}^n$ ,  $\varepsilon > 0$  and  $A \in \mathbb{R}^{n \times p}$ , the sharpness of  $f$  at  $x$ :

$$\phi_{x,f}(\varepsilon, A) := \frac{(\max_{y \in C_\varepsilon} f(x + Ay)) - f(x)}{1 + f(x)} \times 100 \quad (1)$$

- $A$  can be the identity matrix  $I_n$

# Sharpness of Minima

- Sharpness of Minima in Full Space( $A$  is the identity matrix):

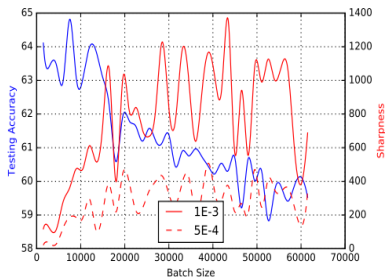
	$\epsilon = 10^{-3}$		$\epsilon = 5 \cdot 10^{-4}$	
	SB	LB	SB	LB
$F_1$	$1.23 \pm 0.83$	$205.14 \pm 69.52$	$0.61 \pm 0.27$	$42.90 \pm 17.14$
$F_2$	$1.39 \pm 0.02$	$310.64 \pm 38.46$	$0.90 \pm 0.05$	$93.15 \pm 6.81$
$C_1$	$28.58 \pm 3.13$	$707.23 \pm 43.04$	$7.08 \pm 0.88$	$227.31 \pm 23.23$
$C_2$	$8.68 \pm 1.32$	$925.32 \pm 38.29$	$2.07 \pm 0.86$	$175.31 \pm 18.28$
$C_3$	$29.85 \pm 5.98$	$258.75 \pm 8.96$	$8.56 \pm 0.99$	$105.11 \pm 13.22$
$C_4$	$12.83 \pm 3.84$	$421.84 \pm 36.97$	$4.07 \pm 0.87$	$109.35 \pm 16.57$

# Outline

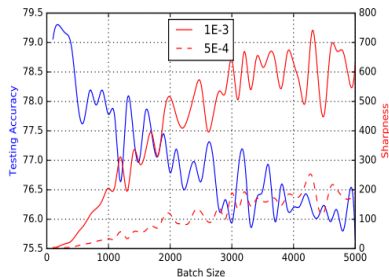
- 1 Introduction
  - Batch Size of Stochastic Gradient Methods
- 2 Drawbacks of Large-Batch Methods
  - Main Observation
  - Numerical Results
  - Parametric Plots
  - Sharpness of Minima
- 3 Success of Small-Batch Methods
  - Deterioration along Increasing of Batch-Size
  - Warm-started Large Batch experiments
- 4 Summary

# Deterioration along Increasing of Batch-Size

- Note batch-size  $\approx 15000$  for  $F_2$  and batch-size  $\approx 500$  for  $C_1$



(a)  $F_2$



(b)  $C_1$

- There exists a threshold after which there is a deterioration in the quality of the model.

# Outline

## 1 Introduction

- Batch Size of Stochastic Gradient Methods

## 2 Drawbacks of Large-Batch Methods

- Main Observation
- Numerical Results
- Parametric Plots
- Sharpness of Minima

## 3 Success of Small-Batch Methods

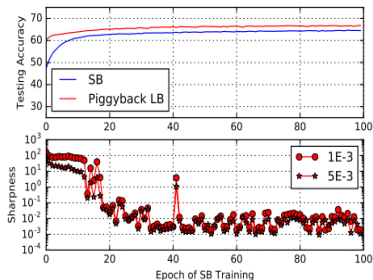
- Deterioration along Increasing of Batch-Size
- Warm-started Large Batch experiments

## 4 Summary

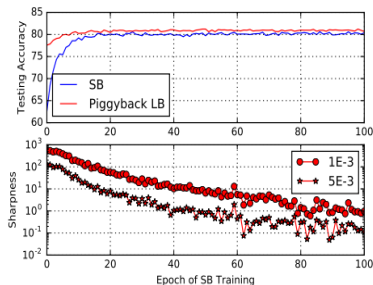


# Warm-started Large Batch experiments

- Train network for 100 epochs with batch-size=256 and use these 100 epochs as starting points.



(a)  $F_2$



(b)  $C_1$

- The SB method needs some epochs to explore and discover a flat minimizer.

# Summary

- Numerical experiments that support the view that convergence to sharp minimizers gives rise to the poor generalization of large-batch methods for deep learning.
- SB methods have an exploration phase followed by convergence to a flat minimizer.
- Attempts to remedy the problem:
  - Data augmentation
  - Conservative training
  - Adversarial training
  - Robust optimization