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

N. Keskar*†, D. Mudigere†, J. Nocedal*, M. Smelyanskiy†, P. T. P. Tang†

*Northwestern University, †Intel

Abstract

1. Observation: when using a larger batch (LB) methods there is a degradation in the quality of the model, as measured by its ability to generalize.
2. We investigate the cause for this generalization drop and present numerical evidence that LB methods tend to converge to sharp minimizers of the training and testing functions.
3. SB methods converge to flat minimizers, and this is due to the inherent noise in the gradient estimation.

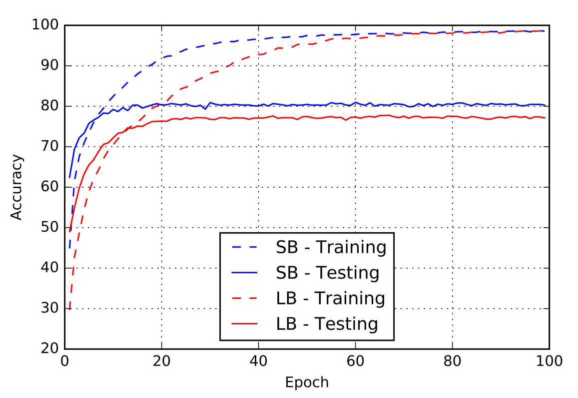
1 Motivation - Why LB Methods ?

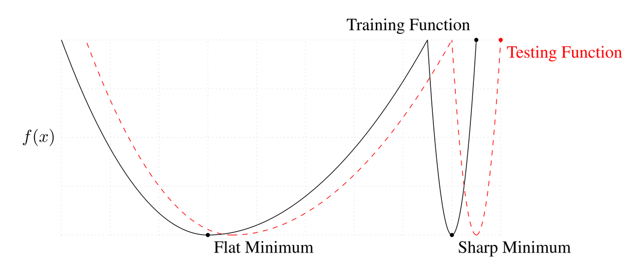
- Small Batch (SB) SGD — Simple, effective but limited scaling (100s of nodes).
- Large Batch (LB) methods — Improved concurrency, potential to scale to 1000+ nodes, faster time-to-train, larger problems.

Generalization Gap with LB methods

	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	$99.10\% \pm 1.23\%$	$99.57\% \pm 1.84\%$	$63.08\% \pm 0.5\%$	$57.81\% \pm 0.17\%$

Conventional Wisdom





- LB methods lack noise/explorative properties.

  "Why is this bad?"
- Minimum number of iterations are required for convergence.

  "Not true; gap exists even if run for 1000s of epochs."
- LB methods "overfit".

  "Once model is specified, unclear what this means. Surely **not** over-training."

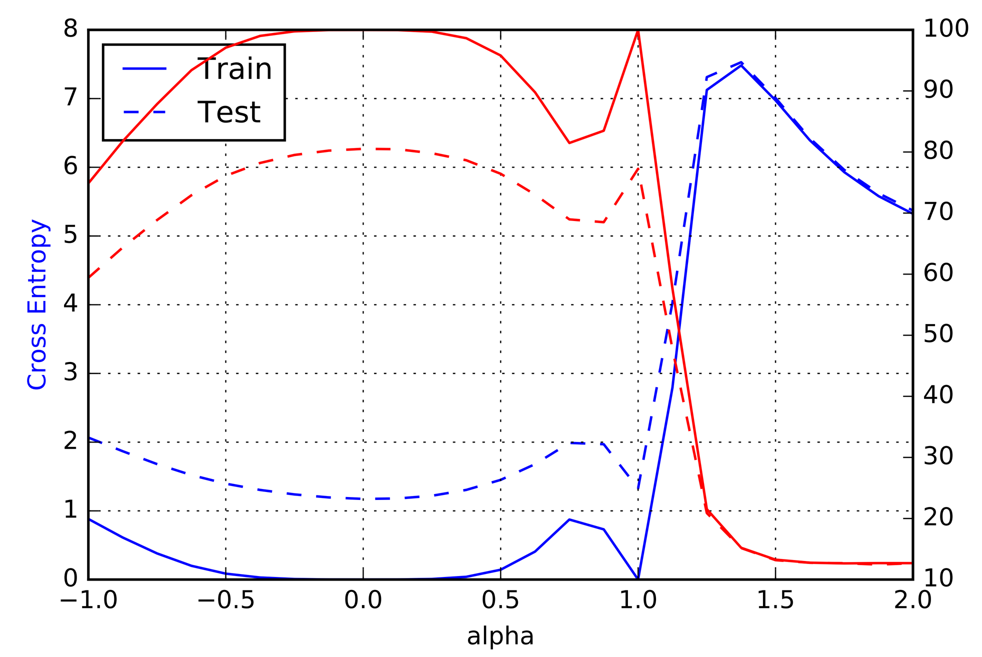
Our Observation

1. The lack of generalization ability is due to the fact that LB methods tend to converge to *sharp minimizers* of the training function.
2. These minimizers are characterized by large positive eigenvalues in $\nabla^2 f(x)$.
3. SB methods converge to flat minimizers characterized by small positive eigenvalues of $\nabla^2 f(x)$.

2 Evidence for Sharpness

Parametric Plots

Plot for $\alpha \in [-1, 2]$: $f(\alpha x_\ell^\star + (1-\alpha)x_s^\star)$



Sharpness Metric

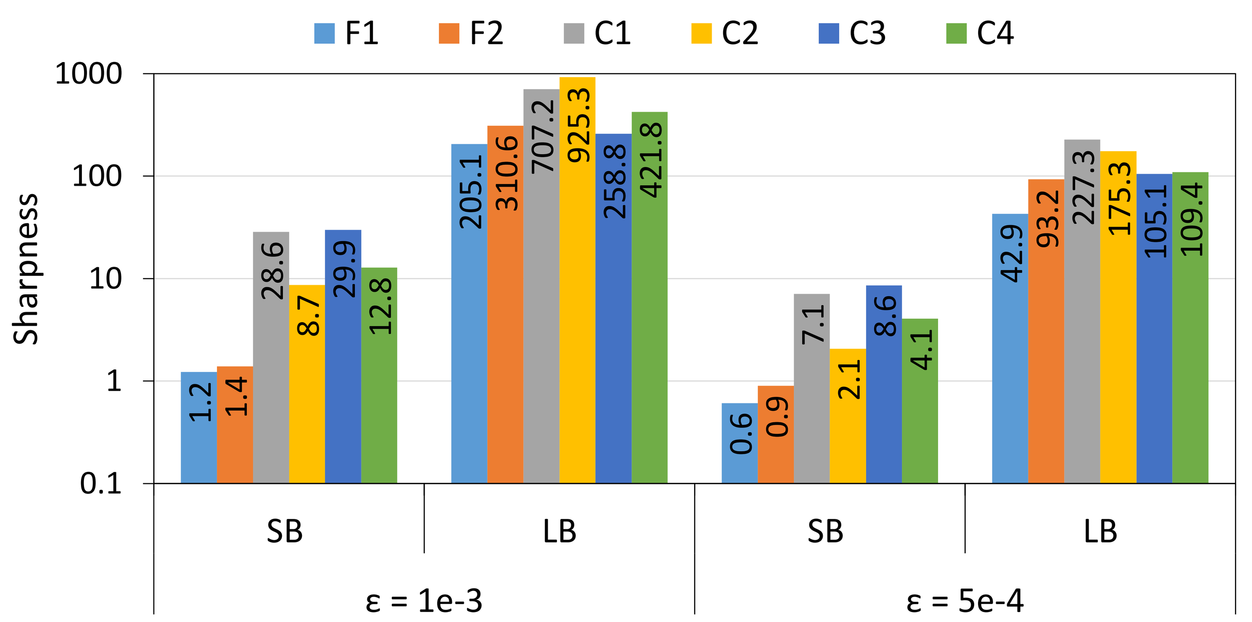
Let $\mathcal{C}_\epsilon$ defined as:

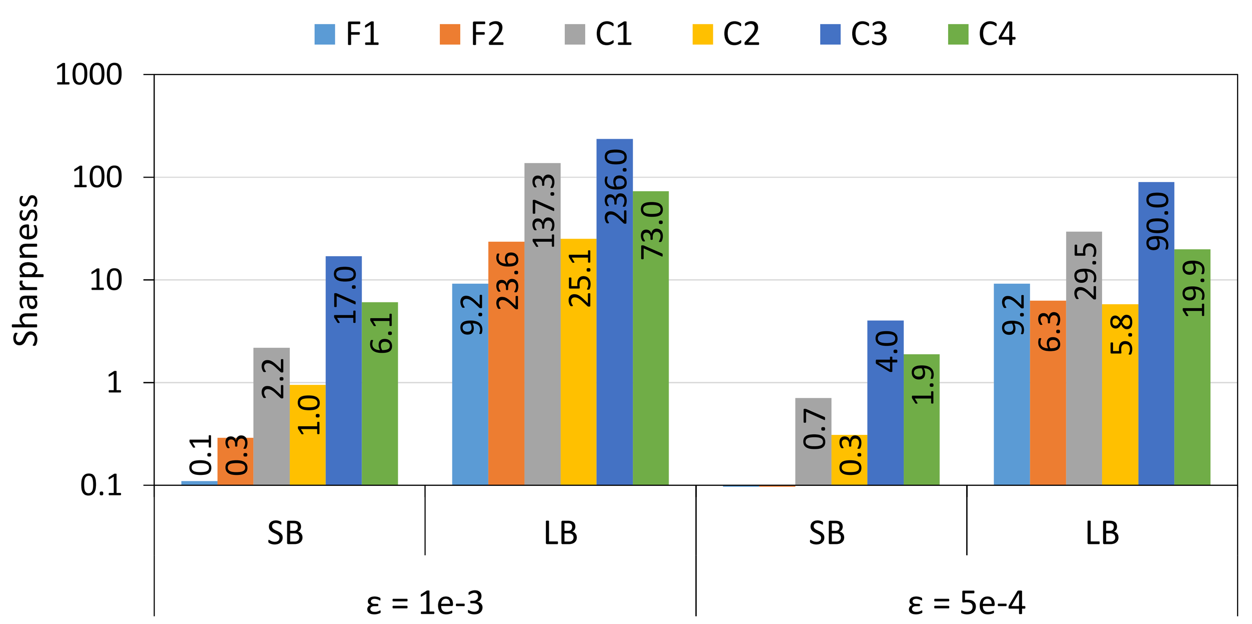
$$\mathcal{C}_\epsilon = \{z \in \mathbb{R}^p : -\epsilon(|(A^+x)_i| + 1) \leq z_i \leq \epsilon(|(A^+x)_i| + 1) \quad \forall i \in \{1, 2, \cdots, p\}\}$$

where A^+ denotes the pseudo-inverse of A.

Metric 2.1. *Given a point* $x \in \mathbb{R}^n$, $\epsilon > 0$ *and* $A \in \mathbb{R}^{n \times p}$, *we define the* $(\mathcal{C}_\epsilon, A)$*-sharpness of* x *as:*

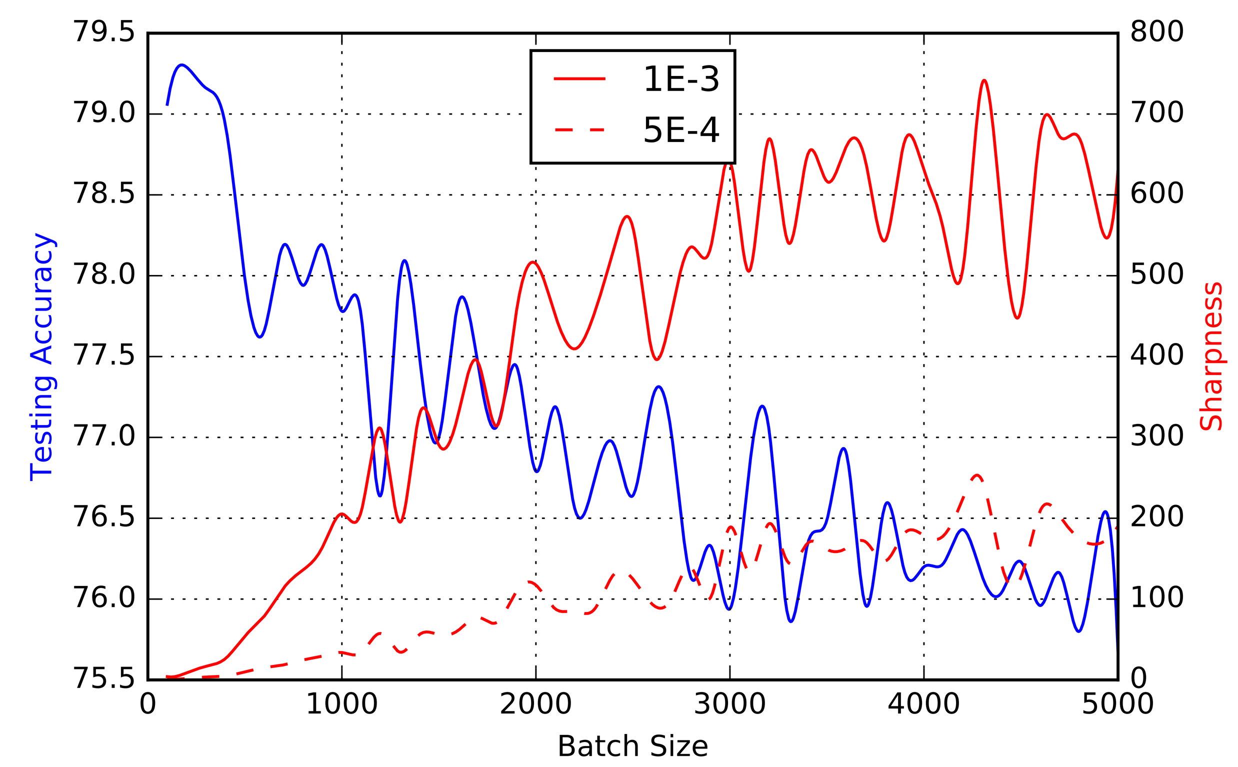
$$\phi_{x,f}(\epsilon, A) := \frac{(\max_{y \in \mathcal{C}_\epsilon} f(x + Ay)) - f(x)}{1 + f(x)} \times 100. \qquad (1)$$



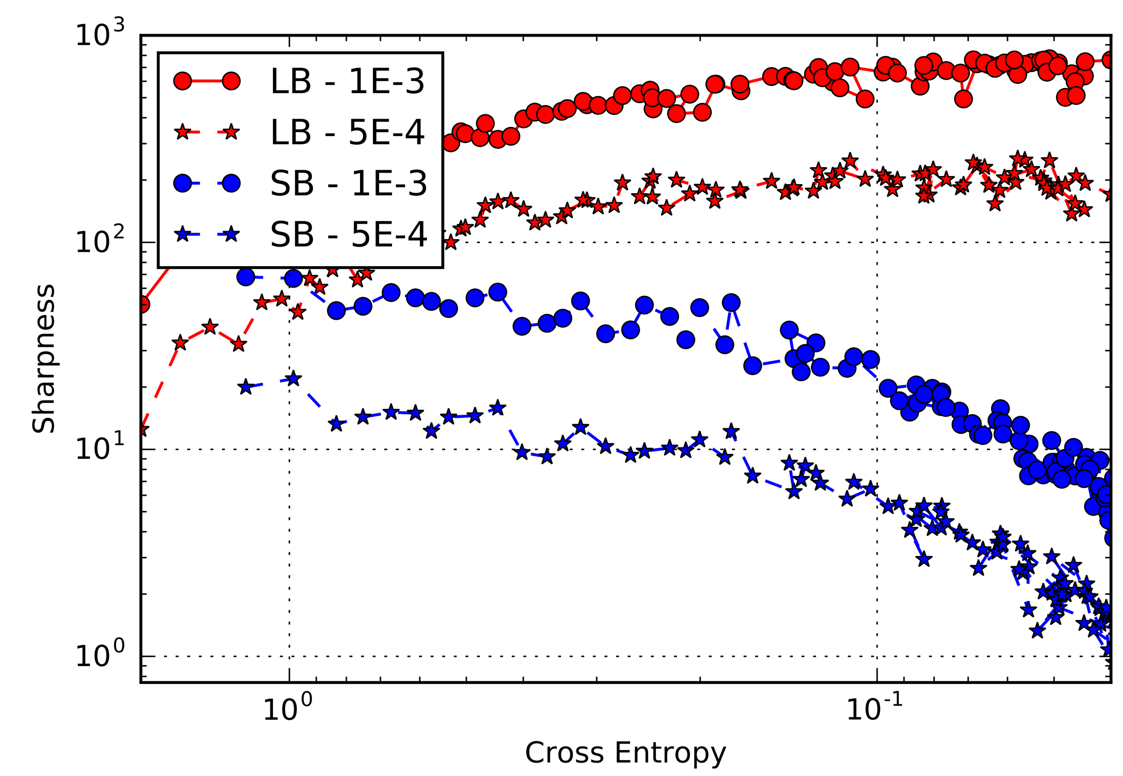


3 Success of SB Methods

Evolution

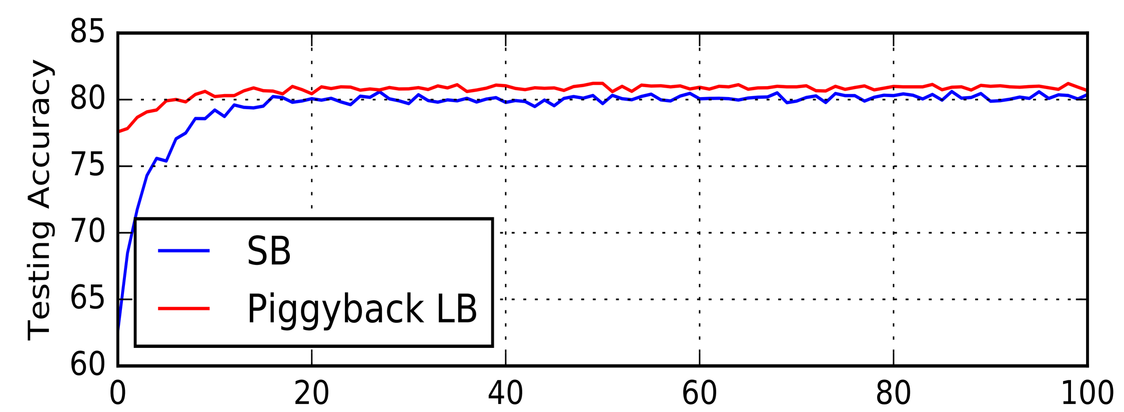


Sharpness v/s Cross Entropy



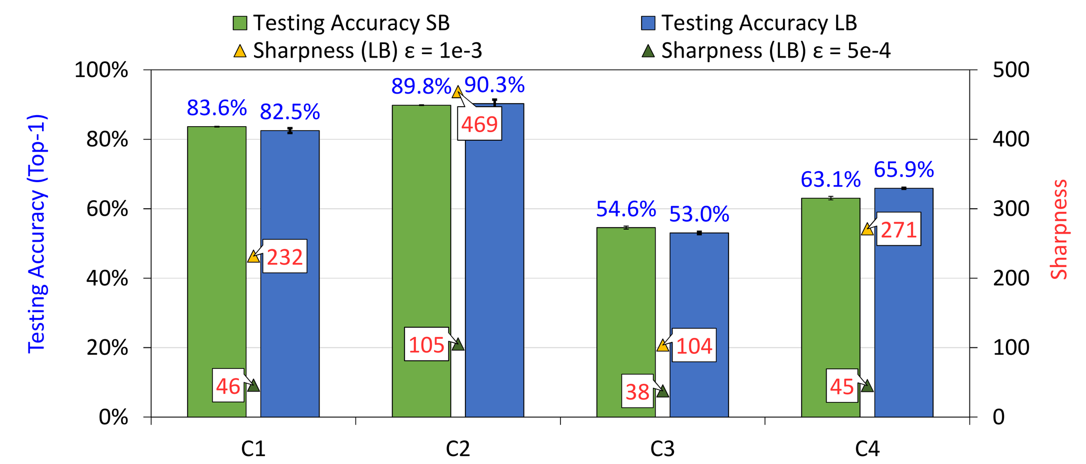
Piggybacking Experiments

- After every epoch of SB method, run 100 epochs of an LB method and plot the testing accuracies.



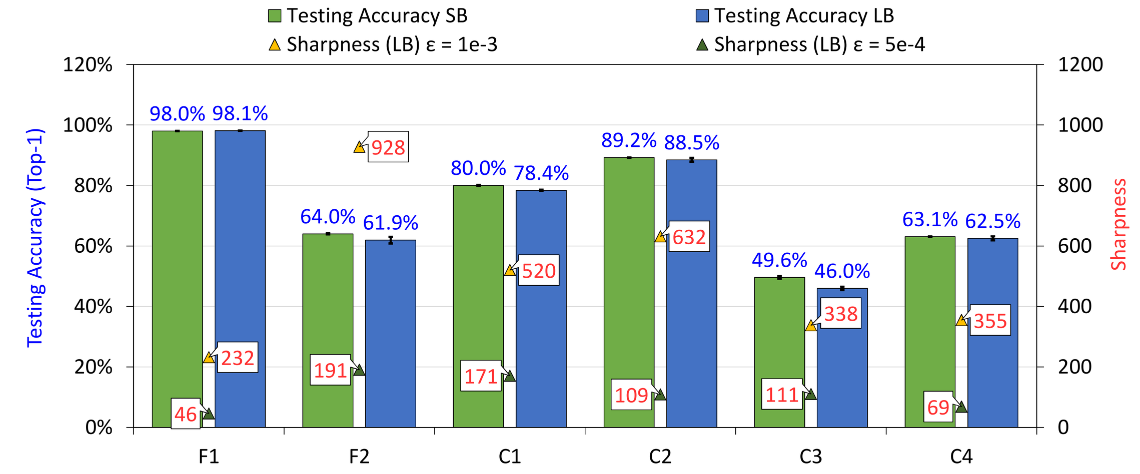
4 Attempts to Improve LB Methods

Data Augmentation



Conservative Training

$$x_{k+1} = \arg\min_x f_{B_k}(x) + \frac{\lambda}{2}\|x - x_k\|_2^2$$



5 Relationship to Recent Work

- **Entropy-SGD**: SGD variant designed to navigate towards flatter minimizers.
- **Sharp Minima *can* Generalize**: (Insufficiency) theoretical analysis of rectifier networks demonstrating weaknesses of sharpness metrics. Decouples training algorithm from generalization argument.
- **Rethinking Generalization**: Classical statistical learning theory cannot explain generalization in deep learning. Explicit regularization is neither necessary nor sufficient; implicit generalization (of SGD) is key.

6 Conclusions

tl; dr

- LB methods → sharp minimizers and these minimizers *correlate* with poorer generalization.
- SB methods avoid sharp minimizers due to noise.
- Our attempts at data augmentation, conservative training, robust optimization, adversarial training etc. did not consistently close the gap.

7 Open Questions

- Rigorous relationship between training algorithm, minimizer properties and generalization performance.
- Steering to flat minimizers; scalable LB method with SOTA performance.