Impact of Mathematical Norms on Convergence of Gradient Descent Algorithms for Deep Neural Networks Learning

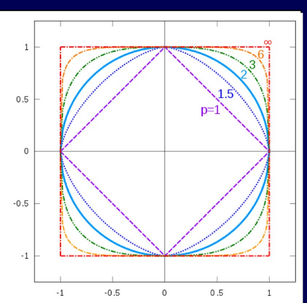
Linzhe Cai ✉, Xinghuo Yu, Chaojie Li, Andrew Eberhard, Lien Thuy Nguyen, and Chuong Thai Doan
s3548838@student.rmit.edu.au

Abstract.

To improve the performance of gradient descent learning algorithms, the impact of different types of norms is studied for deep neural network training. The performance of different norm types used on both finite-time and fixed-time convergence algorithms are compared. The accuracy of the multiclassification task realized by three typical algorithms using different types of norms is given, and the improvement of Jorge's finite time algorithm with momentum or Nesterov accelerated gradient is also studied. Numerical experiments show that the infinity norm can provide better performance in finite time gradient descent algorithms and give strong robustness under different network structures.

Mathematical Norm.

- L_1 *norm* (*Taxicab*): $\|x\|_1 \coloneqq \sum_{i=1}^{n} |x_i|$
- L_2 *norm* (*Euclidean*): $\|x\|_2 \coloneqq \sqrt{{x_1}^2 + {x_2}^2 + \cdots + {x_n}^2}$
- L_p *norm* ($p > 1$): $\|x\|_p \coloneqq \left(\sum_{i=1}^{n} |x_i|^p\right)^{1/p}$
- L_∞*norm* (*Infinity*): $\|x\|_\infty \coloneqq \max_i |x_i|$



Equivalence of Norms.

$$C\,\|x\|_\alpha \le \|x\|_\beta \le D\,\|x\|_\alpha$$

$$\|x\|_p \le \|x\|_r \le n^{\left(\frac{1}{r}-\frac{1}{p}\right)}\,\|x\|_p$$

Finite-Time and Fixed-Time Convergence Algorithms (Euclidean norm based).

Jorge's finite-time:

$$\frac{dw}{dt} = -\frac{\nabla_w J}{\|\nabla_w J\|_2}$$

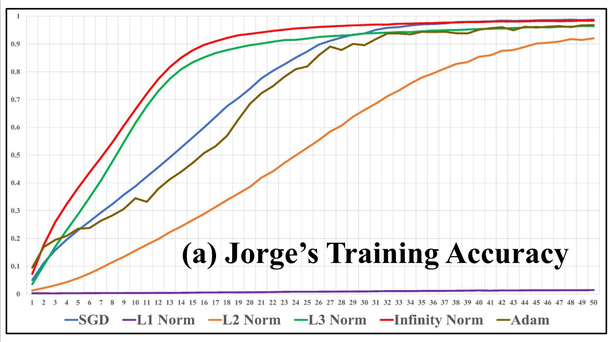
Wibisono's finite-time:

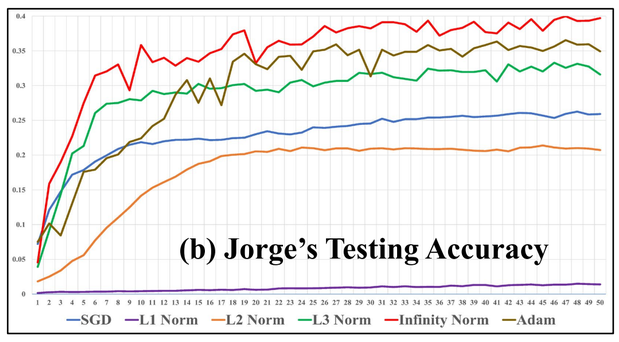
$$\frac{dw}{dt} = -\zeta \frac{\nabla_w J}{\|\nabla_w J\|_2^{\frac{q-1}{q-2}}} \qquad (q > 2)$$

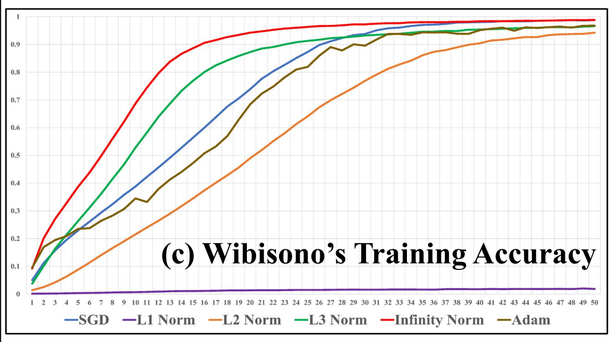
Garg's fixed-time:

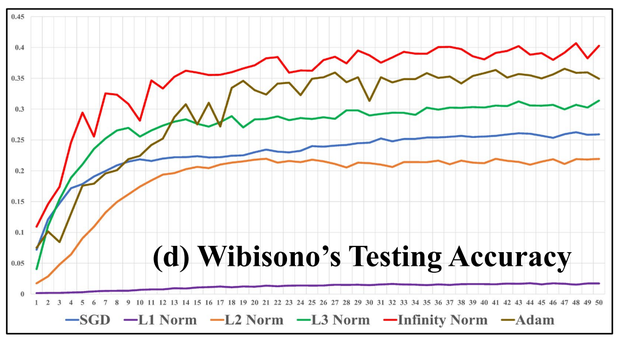
$$\frac{dw}{dt} = -C_1 \frac{\nabla_w J}{\|\nabla_w J\|_2^{\frac{p_1-1}{p_1-2}}} - C_2 \frac{\nabla_w J}{\|\nabla_w J\|_2^{\frac{p_2-1}{p_2-2}}} \qquad (p_1 > 2;\ 1 < p_2 < 2)$$

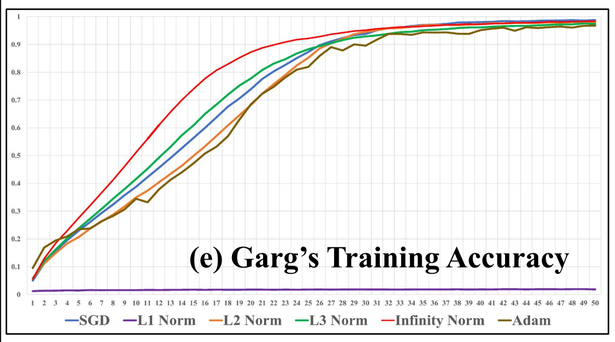
Case Study 1. Three Typical Algorithms Using Different Types of Norms

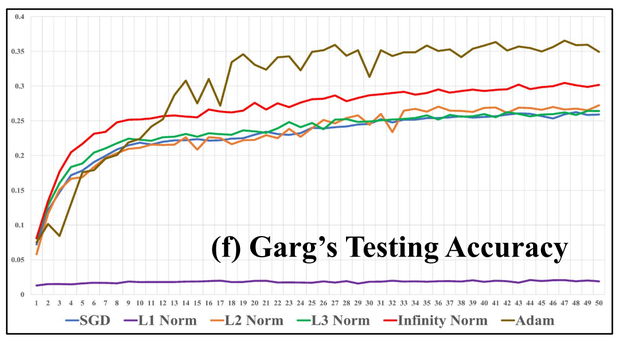










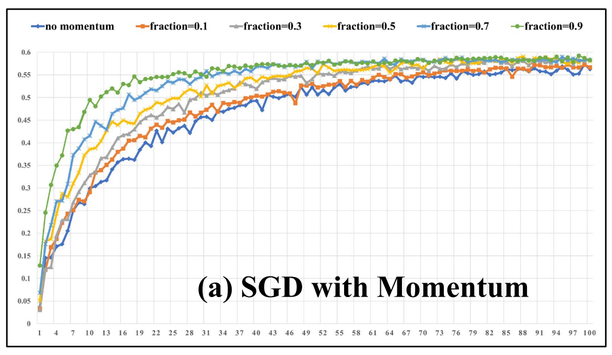


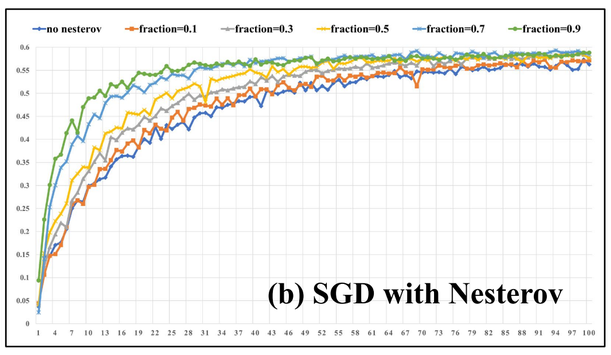
The Performance of three different Algorithms under Different Norms (ResNet50 CIFAR100)

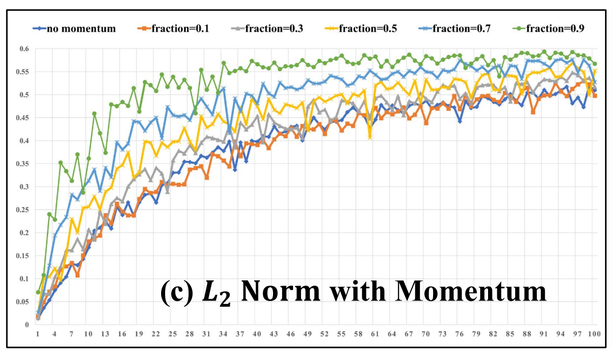
	Statistics	Jorge	Wibisono	Garg	Jorge-v	Wibisono-v	Garg-v
$L_\infty - L_2$	Median	0.3154	0.2383	0.0269	0.1732	0.1687	0.0338
	Mean	0.3209	0.2615	0.0739	0.1721	0.1636	0.0353
$\frac{L_\infty - L_2}{L_2}$	Median	77.92%	46.96%	4.18%	86.99%	82.24%	15.73%
	Mean	169.30%	119.93%	16.64%	126.54%	108.38%	16.18%

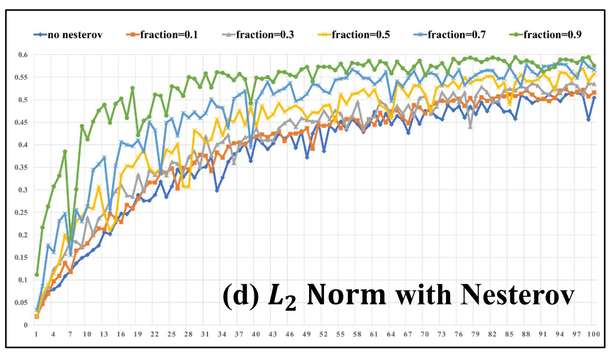
- The effects on different types of norms are obvious for Jorge's finite-time algorithm.
- The performance of Jorge's and Wibisono's Algorithms using Infinity Norm can surpass SGD and Adam for training and testing accuracy during the overall process.

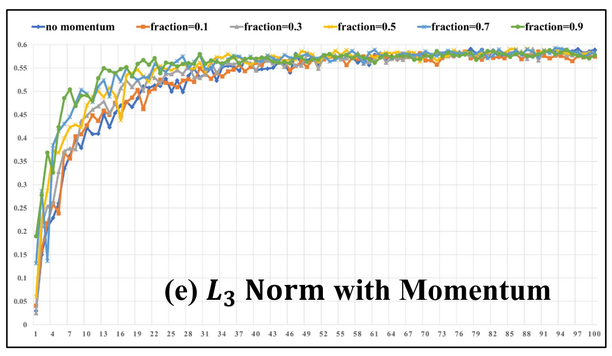
Case Study 2. Jorge's Finite-Time Algorithm with Momentum and Nesterov

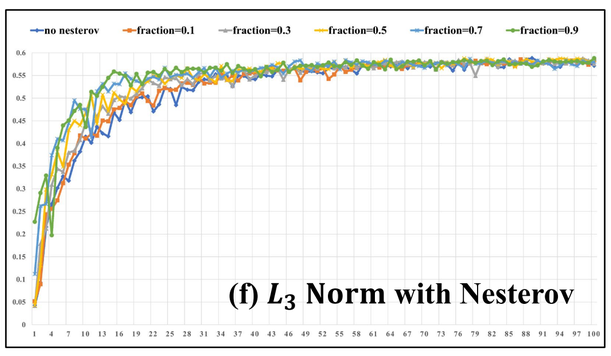


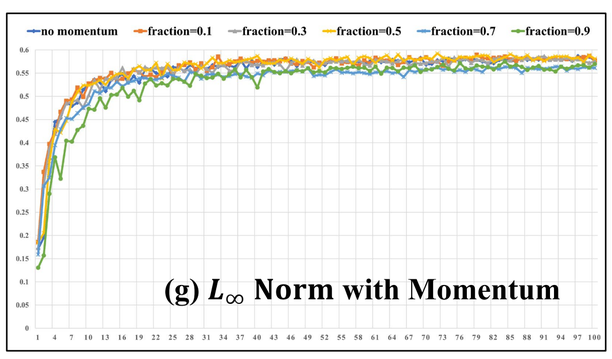


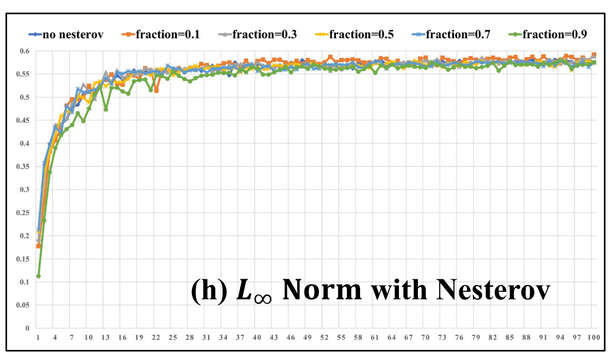








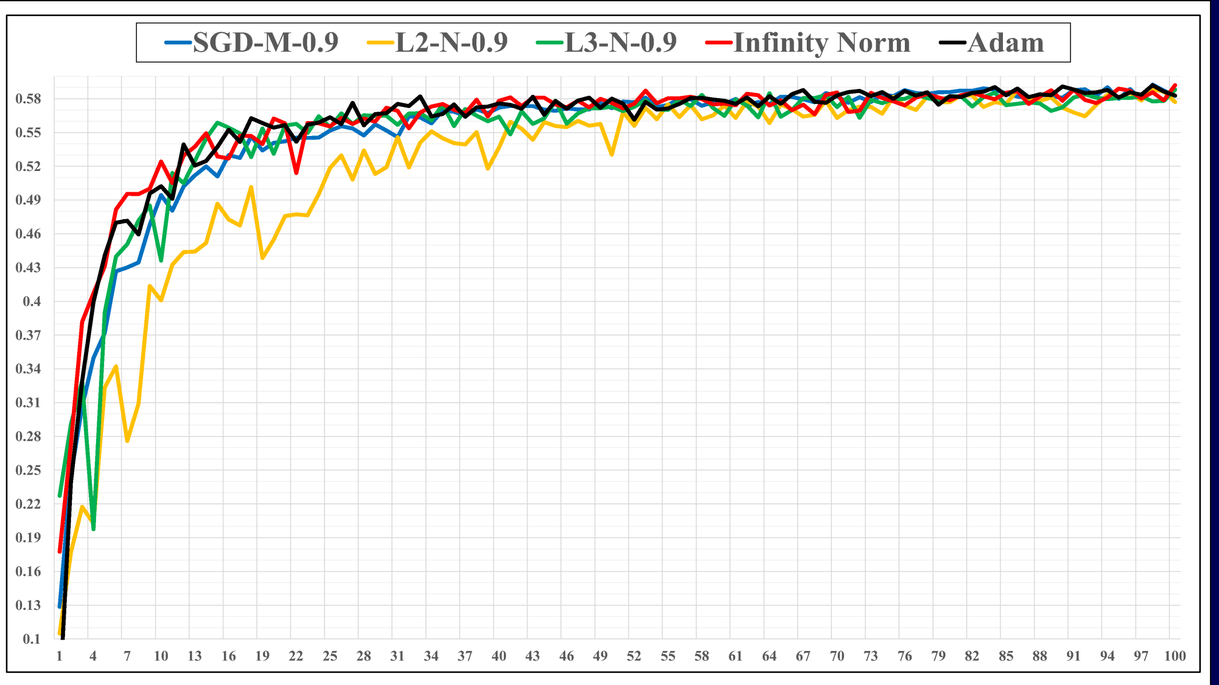




Testing Accuracy of different Norms with the Momentum and Nesterov Accelerate Method (CIFAR100, Six Convolutional Layer Structure)

	Statistics	$L_2 - M$	$L_2 - N$	$L_3 - M$	$L_3 - N$	$L_\infty - M$	$L_\infty - N$
$f_{0.9} - f_0$	Median	0.1349	0.1356	0.0071	0.0134	-0.0254	-0.0143
	Mean	0.1406	0.1499	0.0270	0.0277	-0.0205	-0.0098
$\frac{f_{0.9} - f_0}{f_0}$	Median	31.25%	32.02%	1.27%	2.38%	-5.04%	-3.05%
	Mean	56.46%	62.38%	12.61%	12.03%	-3.57%	-1.72%

- The difference between Momentum and corresponding Nesterov under the same fraction value is not obvious.
- The improvement after involving Momentum and Nesterov is outstanding on the L_2 norm-based Jorge's finite-time algorithm.



Highest CIFAR100 Accuracy using different types of Norms with Momentum or Nesterov

- Even with the help of Momentum or Nesterov Acceleration, SGD and traditional (the L_2 norm-based) Jorge's finite-time algorithm cannot surpass the Adam.
- The infinity norm gradient flow (INGF) without momentum can be imaged as an invisible ceiling among all different types of norms.
- The average running time of Adam is 13.4% longer than which of the INGF for it's adaptive learning rates (computational burden) and history records (memory burden) reasons.

Conclusion: Qualitative analysis after the equivalence of norms with the help of convergence property verifies the convergence rate. The performance of three typical algorithms using different types of norms is quantitatively analysed, and the improvement of Jorge's finite-time algorithm under different types of norms with momentum and Nesterov is studied. Jorge's finite-time algorithm with infinity norm can provide reliable performance.