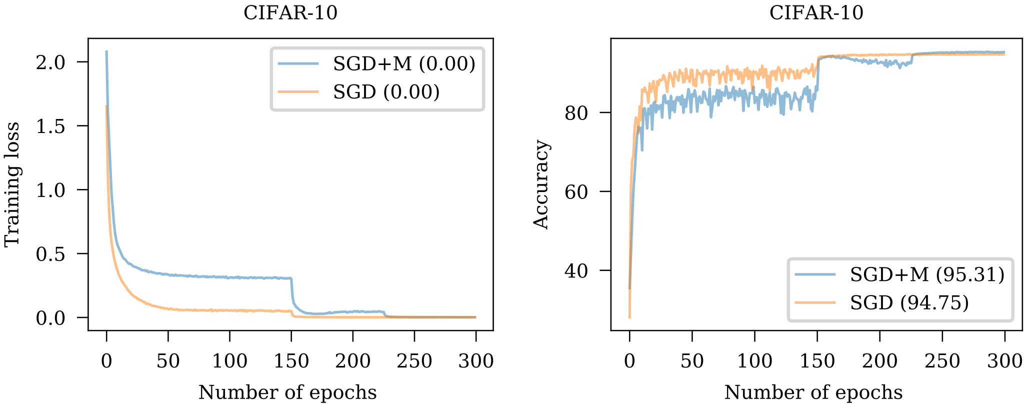
Towards understanding how momentum improves generalization in deep learning

Samy Jelassi$^{(1)}$ and Yuanzhi Li$^{(2)}$

$^{(1)}$Princeton University
$^{(2)}$Carnegie Mellon University

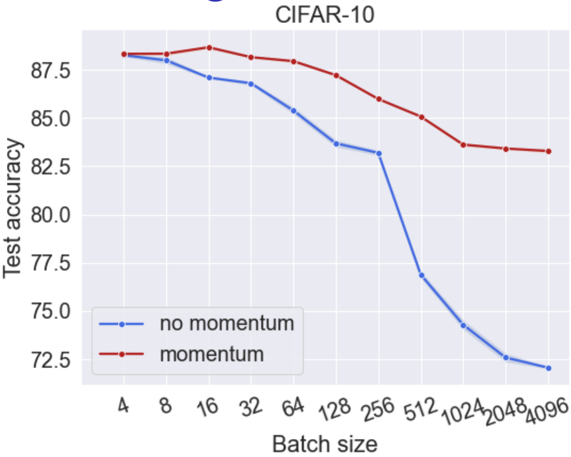
Momentum improves generalization on CIFAR-10



ResNet-18 trained with **data augmentation** and **batch normalization** on CIFAR-10 for 300 epochs. SGD with momentum (SGD+M) gets higher generalization compared to vanilla SGD.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning.

Is the generalization improvement tied to the stochastic noise in the gradient?



VGG-19 trained on CIFAR-10 for 300 epochs. We **turn off** data augmentation and batch normalization. The generalization improvement gets larger as the batch size increases.

Does momentum unconditionally improve generalization in deep learning?

Answer: No ! Binary classification instance with Gaussian data.

Teacher Student	1-MLP	2-MLP	1-CNN	2-CNN
1-MLP	1.00	1.00	1.00	0.99
2-MLP	0.99	1.00	0.99	0.99
1-CNN	0.99	1.00	1.00	1.01
2-CNN	1.00	1.00	1.00	1.02

Ratio Test(GD+M)/Test(GD) when training ReLU networks on **Gaussian** synthetic dataset. Training for 1000 epochs to ensure tiny training error. Averaged over 3 runs.

Main message of this talk

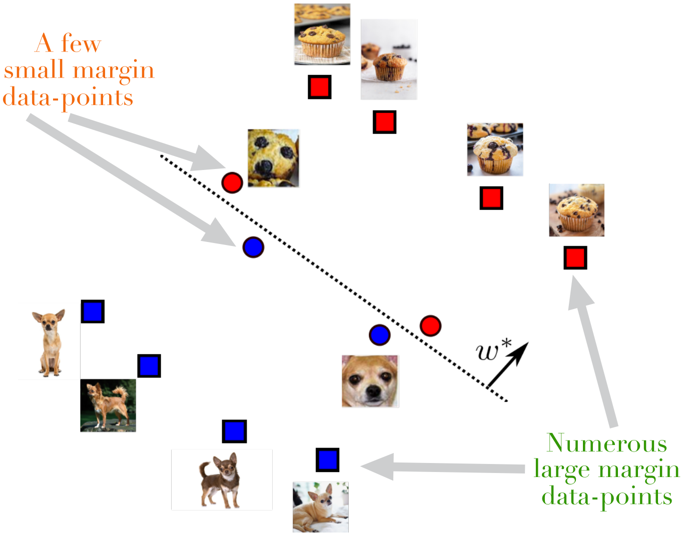
The generalization improvement induced by momentum in deep learning is not tied to the stochastic noise of the gradient but rather depends on both the structure of the data and the learning problem.

Main message of this talk

The generalization improvement induced by momentum in deep learning is not tied to the stochastic noise of the gradient but rather depends on both the structure of the data and the learning problem.

Contribution: We construct a binary classification problem where GD+M provably outperforms GD in terms of generalization. Such improvement cannot be obtained by tuning the learning rate in GD.

Our binary classification dataset



Theorem

There exists an over-parametrized two-layer CNN such that trained on our binary classification dataset:

	Training loss	Test accuracy (large margin data)	Test accuracy (small margin data)
Gradient Descent	✓	✓	✗
Gradient Descent + momentum	✓	✓	✓

Theorem

There exists an over-parametrized two-layer CNN such that trained on our binary classification dataset:

	Training loss	Test accuracy (large margin data)	Test accuracy (small margin data)
Gradient Descent	✓	✓	✗
Gradient Descent + momentum	✓	✓	✓

Key insight: Historical gradients in momentum gradient help to learn small margin data.

Theorem

There exists an over-parametrized two-layer CNN such that trained on our binary classification dataset:

	Training loss	Test accuracy (large margin data)	Test accuracy (small margin data)
Gradient Descent	✓	✓	✗
Gradient Descent + momentum	✓	✓	✓

Poster: Hall E #237