Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence

Challenge: How to tune the hyper-parameters of SGD to make deep learning generalize well?

Theoretical analysis: We analyse the generalizationability of SGD via stochastic differential equation:

- Model the updates of SGD as a Ornstein-Uhlenbeck process;

$$\Delta\theta(t) = \theta(t+1) - \theta(t) = -\eta g(\theta) + \frac{\eta}{|S|} B \Delta W, \Delta W \sim \mathcal{N}(0, I),$$

where $\theta(t)$ is the weight in time (step) t, η is the step size, $|S|$ is the batch size.

- Use the stationary distribution to express the output of SGD;

$$q(\theta) = M \exp\left\{-\frac{1}{2}\theta^\top \Sigma^{-1} \theta\right\},$$

where

$$\Sigma A + A\Sigma = \frac{\eta}{|S|} B B^\top,$$

and A expresses the local geometry around the global minima:

$$\mathcal{R}(\theta) = \frac{1}{2}\theta^\top A \theta.$$

- Finally, we get a PAC-Bayesian generalization bound for SGD:

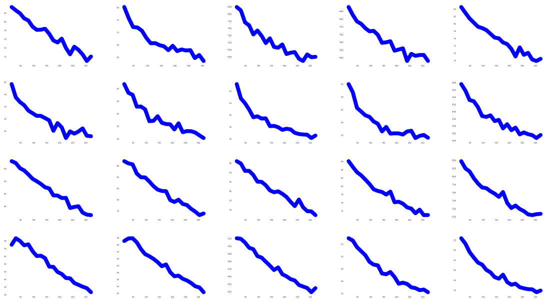
$$R(Q) \leq \hat{R}(Q) + \sqrt{\frac{d\log\left(\frac{|S|}{\eta}\right) + \log\left(\frac{\det(A)}{\det(B)^2}\right) + \frac{1}{2}tr(\Sigma - I) + 2\log\frac{1}{\delta} + 5\log N + 17}{4N-2}},$$

Results: The generalization ability of SGD has

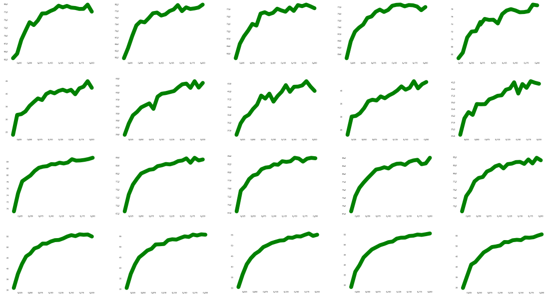
- A negative correlation with batch size;
- A positive correlation with learning rate;
- A negative correlation with the ratio of batch size to learning rate.

Empirical analysis: We trained around 1,600 models based on the architectures ResNet-19 and VGG-110 on the datasets CIFAR-10 and CIFAR-100. The results fully support the theoretical results.

Test accuracy vs. batch size*

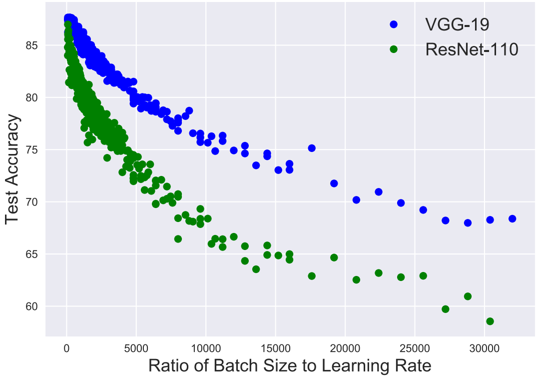


Test accuracy vs. learning rate*

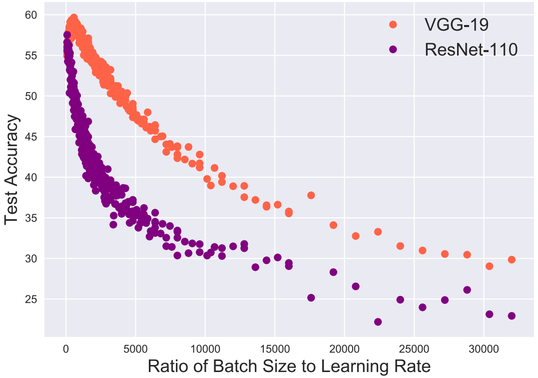


*Every curve is drawn based on the basis that strictive controls irrelevant variables. From top to bottom, the four lines are respectively (1) ResNet-110 on CIFAR-10, (2) ResNet-110 on CIFAR-100, (3) VGG-19 on CIFAR-10, and (4) VGG-19 on CIFAR-10.

Plot of ResNet-110 and VGG-19 on CIFAR-10



Plot of ResNet-110 and VGG-19 on CIFAR-100



He, Liu, and Tao, “Control Batch Size and Learning Rate to Generalize Well: Theoretical and Empirical Evidence”, NeurIPS, 2019.