TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Stochastic Gradient Descent (SGD)

Decoupling the Learning Rate From the Batch Size

Decoupling η from B

For vanilla SGD with minibatching we have

$$\Phi_{t+1} \texttt{ -= } \eta \hat{g}_t$$

$$\hat{g}_t = \frac{1}{B} \sum_b \hat{g}_{t,b}$$

Where $\hat{g}_{t,b}$ is the gradient of the element b of the batch.

Decoupling η from B

For batch size 1 on the same sequence of data points with $b \in \{1, \ldots, B\}$ and with learning rate η_0 we have

$$\begin{aligned} \Phi_{t+B} &= \Phi_t - \sum_b \eta_0 \, \nabla_\Phi \mathcal{L}(b, \Phi_{t+b-1}) \\ &\approx \eta_0 \sum_b \nabla_\Phi \mathcal{L}(b, \Phi_t) \\ &= B\eta_0 \, \hat{g}_t \end{aligned}$$

If η_0 is the optimal learning rate for $B = 1$ then $B\eta_0$ should be the optimal learning rate for general B.

Decoupling η from B

Recent work has show that using $\eta = B\eta_0$ leads to effective learning with very large (highly parallel) batches.

Accurate, Large Minibatch SGD: Training ImageNet in 1 Hour, Goyal et al., 2017.

END