TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Stochastic Gradient Descent (SGD)

The Learning Rate as Temperature

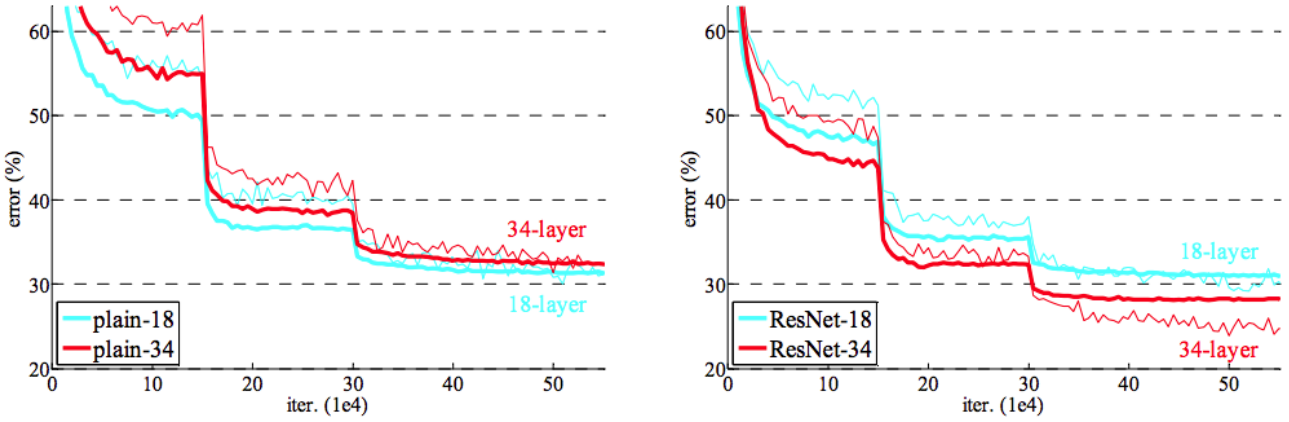
Learning Rate as Temperature

The learning rate can be interpreted as a temperture.

If we run for a long time at a large learning rate we converge to a noisy (hot) stationary distribution with a high loss value.

At a lower learning rate we converge to a cooler stationary distribution with a lower loss value.

Learning Rate as Temperature



These Plots are from the original ResNet paper. Left plot is for CNNs without residual skip connections, the right plot is ResNet.

Thin lines are training error, thick lines are validation error.

In all cases η is reduced twice, each time by a factor of 2.

Batch Size and Temperature

Vanilla SGD with minibatching typically uses the following update which defines the meaning of η.

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

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

Here $\hat{g}_b$ is average gradient over the batch.

Under this update **increasing the batch size (while holding η fixed) reduces the temperature.**

Making Temperature Independent of B

For batch size 1 with learning rate η_0 we have

$$\Phi_{t+1} = \Phi_t - \eta_0 \, \nabla_\Phi \mathcal{L}(t, \Phi_t)$$

$$\begin{aligned} \Phi_{t+B} &= \Phi_t - \sum_{b=0}^{B-1} \eta_0 \, \nabla_\Phi \mathcal{L}(t+b, \Phi_{t+b-1}) \\ &\approx \Phi_t - \eta_0 \sum_b \nabla_\Phi \mathcal{L}(t+b, \Phi_t) \\ &= \Phi_t - B\eta_0 \, \hat{g}_t \end{aligned}$$

For batch updates $\Phi_{t+1} = \Phi_t - B\eta_0 \; \hat{g}_t$ the temperature is essentially determined by η_0 independent of B.

Making Temperature Independent of 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