TTIC 31230, Fundamentals of Deep Learning

David McAllester, Autumn 2020

Stochastic Gradient Descent (SGD)

The Learning Rate as Temperature

Temperature

Physical temperature is a relationship between the energy and probability.

$$P(x) = \frac{1}{Z}\, e^{\frac{-E(x)}{kT}} \qquad Z = \sum_x e^{\frac{-E(x)}{kT}}$$

This is called the Gibbs or Boltzman distribution.

$E(x)$ is the energy of physical microstate state x.

k is Boltzman's constant.

Z is called the partition function.

Temperature

Boltzman's constant can be measured using the ideal gas law.

$$pV = NkT$$

$$\begin{aligned} p &= \text{pressure} \\ V &= \text{volume} \\ N &= \text{the number of molecules} \\ T &= \text{temperature} \\ k &= \text{Boltzman's constant} \end{aligned}$$

We can measure p, V, N and T and solve for k.

Temperature

The Gibbs distribution is typically written as

$$P(x) = \frac{1}{Z}\, e^{-\beta E(x)}$$

$\beta = \frac{1}{kT}$ is the (inverse) temperature parameter.

"Hot" is when β is small and "cold" is when β is large (confusing).

Temperature

In a softmax with a temperature parameter we replace energy $E(x)$ with a score $s(x)$ and drop the negative sign.

$$\underset{y}{\text{softmax}} = \frac{1}{Z}\, e^{\beta s(y)}$$

We can think of the temperature parameter β as simply a parameter of this distribution.

Learning Rate as Temperature

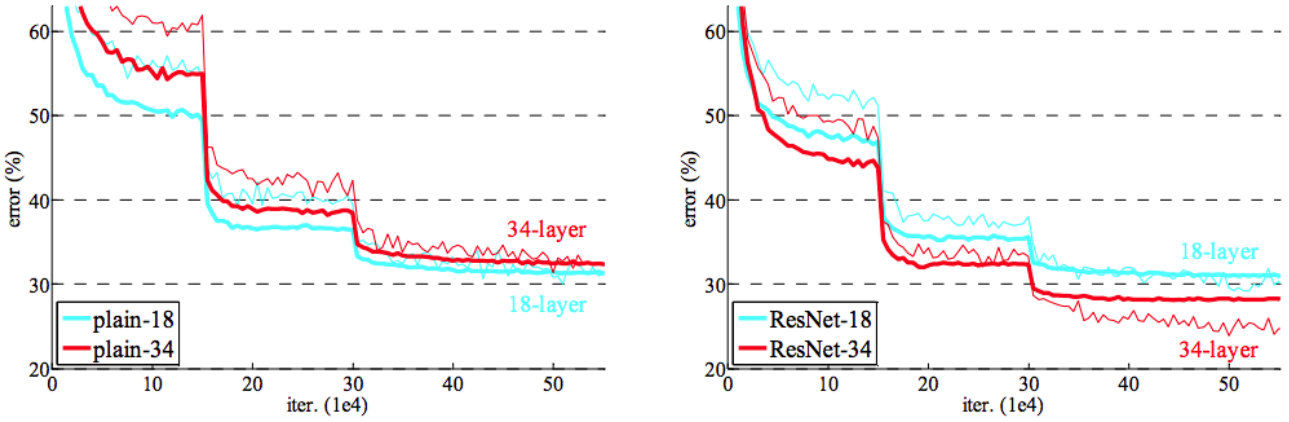
A finite learning rate defines an equalibrium probability distribution (or density) over the model parameters.

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

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

We will later give a stochastic differential equation (SDE) model for SGD and derive a formal relationship between the learning rate and the temperature parameter of the Boltzman distribution.

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 the 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

In 2017 it was discovered that setting $\eta = B\eta_0$ allows very large (highly parallel) batches.

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

END