TTIC 31230, Fundamentals of Deep Learning

David McAllester, Winter 2020

Stochastic Gradient Descent (SGD)

Gradient Flow

Gradient Flow

Gradient flow is a non-stochastic (deterministic) model of stochastic gradient descent (SGD).

Gradient flow is defined by the total gradient differential equation

$$\frac{d\Phi}{dt} = -g(\Phi) \qquad g(\Phi) = \nabla_\Phi \; E_{(x,y)\sim\mathrm{Train}} \; \mathcal{L}(\Phi, x, y)$$

We let $\Phi(t)$ be the solution to this differential equation satisfying $\Phi(0) = \Phi_{\mathrm{init}}$.

Gradient Flow

$$\frac{d\Phi}{dt} = -g(\Phi)$$

For small values of Δt this differential equation can be approximated by

$$\Delta\Phi = -g(\Phi)\Delta t$$

Time as the Sum of the Learning Rates

Consider the total SGD update.

$$\Delta\Phi = -g\Delta t$$

Here Δt has both a natural interpretation as time in a numerical simulation of the flow differential equation.

But it also has a natural interpretation as a learning rate.

This leads to interpreting the sum of the learning rates as "time" in SGD.

Gradient Flow and SGD

Consider a sequence of model parameters Φ_1, …, Φ_N produced by SGD with

$$\Phi_{i+1} = \Phi_i - \eta \hat{g}_i$$

and where $\hat{g}_i$ is the gradient of the ith randomly selected training point.

Take $\eta \rightarrow 0$ and $N \rightarrow \infty$ using $N = t/\eta$. We will show that in this limit for SGD we have that Φ_N converges to $\Phi(t)$ as defined by gradient flow.

Gradient Flow and SGD

For $\Phi_{i+1} = \Phi_i - \eta \hat{g}_i$ we divide Φ_1, …, Φ_N into $\sqrt{N}$ blocks.
$(\Phi_1, \ldots, \Phi_{\sqrt{N}})$ $(\Phi_{\sqrt{N}+1}, \ldots, \Phi_{2\sqrt{N}})$ $\cdots$ $(\Phi_{T-\sqrt{N}+1}, \ldots, \Phi_N)$

For $\eta \to 0$ and $N = t/\eta$ we have $\eta\sqrt{N} \to 0$ which implies

$$\Phi_{\sqrt{N}} \sim \Phi_0 - \eta\sqrt{N} g$$

where g is the average (non-stochastic) gradient.

Since the gradients within each block become non-stochastic, we are back to gradient flow.

END