RMS Prop. A very effective, but currently unpublished adaptive learning rate method.

Instead of using $\frac{\eta}{\sqrt{G_{i,i}^{(t)} + \varepsilon}}$ as the learning rate, we decay the historical gradients:

$$(*) \quad G_{i,i}^{(t+1)} \leftarrow \underline{\text{decay\_rate}} * G_{i,i}^{(t)} + \underline{(1 - \text{decay\_rate})} * g_i^{(t)^2}$$

Adam. A recently proposed update that looks a bit like RMSprop with momentum.

In addition to using equation $(*)$ above to adjust the learning rate, we also keep track of the momentum:

$$m^{(t+1)} = \beta_1 \cdot m^{(t)} + (1 - \beta_1) \cdot \frac{\partial f}{\partial w} \text{ (or } g_t^{(t)})$$

$$v^{(t+1)} = \beta_2 \cdot v^{(t)} + (1 - \beta_2) \cdot (g_t^{(t)})^2$$

It turns out that the decaying terms above biases the gradients towards zero during initial training (recall the learning slowdown problem). Therefore, Adam counteracts by correcting the bias as follows:

$$\hat{m}^{(t+1)} = \frac{m^{(t+1)}}{1 - \beta_1^{t+1}}$$

$$\hat{v}^{(t+1)} = \frac{v^{(t+1)}}{1 - \beta_2^{t+1}}$$

Summary. Momentum, Adagrad, RMSprop, and Adam: These methods may seem mysterious if this is the first time you learn about them. My understanding is that these methods should be viewed as rules of thumb that have been tested by previous researchers. They represent an accumulation of experience by deep learning researchers.

Nesterov accelerated gradient,

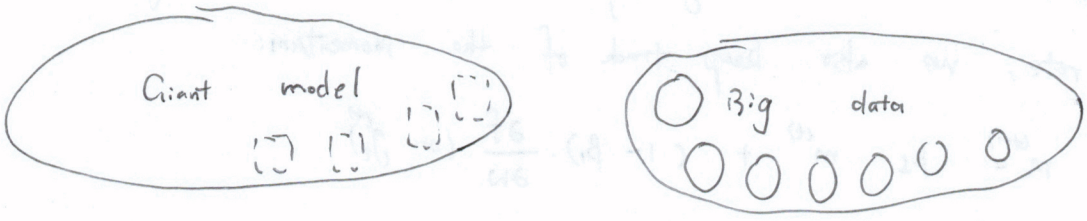
Parallel / Distributed SGD

Given the ubiquity of large-scale data solutions and the availability of low-commodity servers, distributing SGD to speed it up further is an obvious choice.

SGD by itself is inherently sequential. Tradeoff:

Running SGD asynchronously is faster, but suboptimal performance between workers can lead to poor convergence.

Hogwild!



Idea: each mini-batch SGD only updates a small portion of the model parameters!

Mini-batch SGD

Curriculum learning

often, SGD is applied over a random permutation of the training dataset. In other words, each sample is treated as interchangably w/ the other samples.

Idea: Supplying the training examples in a meaningful order may actually lead to improved performance and better convergence.

Conclusions

- Weight initialization: large v.s. small random initialization
- Variants of SGD:
  - Momentum, NAG
  - learning rate: Adagrad, RMSprop, Adam
  - Advanced techniques, Hogwild! Curriculum Learning.