TTIC 31230, Fundamentals of Deep Learning

David McAllester, April 2017

An SGD Progress Theorem

Some Theory

We will prove that minibatch SGD for a **sufficiently large batch size** (for gradient estimation) and a **sufficient small learning rate** (to avoid gradient drift) is guaranteed (with high probability) to reduce the loss.

This guarantee has two main requirements.

- A smoothness condition to limit gradient drift.

- A bound on the gradient norm allowing high confidence gradient estimation.

Smoothness: The Hessian

We can make a second order approximation to the loss.

$$\ell(\Theta + \Delta\Theta) \approx \ell(\Theta) + g^\top \Delta\Theta + \frac{1}{2}\Delta\Theta^\top H \Delta\Theta$$

$$g = \nabla_\Theta\, \ell(\Theta)$$

$$H = \nabla_\Theta \nabla_\Theta\, \ell(\Theta)$$

here H is the second derivative of ℓ, the Hessian matrix.

$$H_{i,j} = \frac{\partial^2 \ell(\Theta)}{\partial\Theta_i\, \partial\Theta_j}$$

The Smoothness Condition

We will assume

$$||H\Delta\Theta|| \leq L||\Delta\Theta||$$

We now have

$$\Delta\Theta^\top H\Delta\Theta \leq L||\Delta\Theta||^2$$

Using the second order mean value theorem one can prove

$$\ell(\Theta + \Delta\Theta) \; \leq \; \ell(\Theta) + g^\top\Delta\Theta + \frac{1}{2}L||\Delta\Theta||^2$$

A Concentration Inequality for Gradient Estimation

Consider a vector mean estimator where the vectors g_n are drawn IID.

$$g_n = \nabla_\Theta \ell_n(\Theta) \qquad \hat{g} = \frac{1}{k}\sum_{n=1}^{k} g_n \qquad g = \mathrm{E}_n\left[\nabla_\Theta\, \ell_n(\Theta)\right]$$

If with probability 1 over the draw of n we have $|(g_n)_i - g_i| \leq b$ for all i then with probability of at least $1-\delta$ over the draw of the sample

$$||\hat{g} - g|| \leq \frac{\gamma}{\sqrt{k}} \qquad \gamma = b\left(1 + \sqrt{2\ln(1/\delta)}\right)$$

Norkin and Wets "Law of Small Numbers as Concentration Inequalities ...", 2012, theorem 3.1

$$
\begin{aligned}
\ell(\Theta + \Delta\Theta) &\leq \ell(\Theta) + g^\top \Delta\Theta + \frac{1}{2} L ||\Delta\Theta||^2 \\
\ell(\Theta - \eta\widehat{g}) &\leq \ell(\Theta) - \eta g^\top \widehat{g} + \frac{1}{2} L \eta^2 ||\widehat{g}||^2 \\
&= \ell(\Theta) - \eta(\widehat{g} - (\widehat{g} - g))^\top \widehat{g} + \frac{1}{2} L \eta^2 ||\widehat{g}||^2 \\
&= \ell(\Theta) - \eta||\widehat{g}||^2 + \eta(\widehat{g} - g)^\top \widehat{g} + \frac{1}{2} L \eta^2 ||\widehat{g}||^2 \\
&\leq \ell(\Theta) - \eta||\widehat{g}||^2 + \eta\frac{\gamma}{\sqrt{k}}||\widehat{g}|| + \frac{1}{2} L \eta^2 ||\widehat{g}||^2 \\
&= \ell(\Theta) - \eta||\widehat{g}|| \left(||\widehat{g}|| - \frac{\gamma}{\sqrt{k}} \right) + \frac{1}{2} L \eta^2 ||\widehat{g}||^2
\end{aligned}
$$

Optimizing η

Optimizing η we get

$$||\widehat{g}|| \left(||\widehat{g}|| - \frac{\gamma}{\sqrt{k}} \right) = -L\eta ||\widehat{g}||^2$$

$$\eta = \frac{1}{L}\left(1 - \frac{\gamma}{||\widehat{g}||\sqrt{k}}\right)$$

Inserting this into the guarantee gives

$$\ell(\Theta - \eta\widehat{g}) \leq \ell(\Theta) - \frac{L}{2}\eta^2 ||\widehat{g}||^2$$

Optimizing k

Optimizing progress per sample, or maximizing η^2/k, we can optimize for k as follows.

$$
\begin{aligned}
\frac{\eta^2}{k} &= \frac{1}{L^2}\left(\frac{1}{\sqrt{k}} - \frac{\gamma}{||\widehat{g}||k}\right)^2 \\
0 &= -\frac{1}{2}k^{-\frac{3}{2}} + \frac{\gamma}{||\widehat{g}||}k^{-2} \\
k &= \left(\frac{2\gamma}{||\widehat{g}||}\right)^2 \\
\eta &= \frac{1}{2L}
\end{aligned}
$$

Warning

If SGD is best viewed as a kind of MCMC (performing exploration) then we probably do not want to guarantee progress.

END