ϵ Convergence Of Gradient Descent With δ Close Gradient Estimates

Dhruv Malik

March 14th 2018

The setup is as follows. Consider a function $f(x)$ which has L-Lipschitz gradients:

$$f(y) \leq f(x) + \nabla f(x)^\top (y - x) + \frac{L}{2} \|y - x\|_2^2 \tag{1}$$

and also satisfies the PL Inequality:

$$\frac{1}{2} \|\nabla f(x)\|_2^2 \geq \mu(f(x) - f(x^*)) \tag{2}$$

This proof shows a notion of convergence with gradient descent, where we are only given δ good estimates of the gradients at each step.

Theorem. *Assume that at each step we are given oracle access to $g(x)$, where $\|g(x) - \nabla f(x)\|_2 < \delta$. Assume that we estimated the gradients well enough that $\delta < \sqrt{\mu\epsilon}$. We follow the update rule:*

$$x_{k+1} = x_k - \frac{1}{L} g(x_k) \tag{3}$$

If we perform N iterations of gradient descent where:

$$N > \frac{2L}{\mu} \log(\frac{f(x_0) - f(x^*)}{\epsilon}) \tag{4}$$

then we have that:

$$\min_{i \in \{1 \ldots N\}} f(x_i) - f(x^*) < \epsilon \tag{5}$$

Proof. By our assumption, we have that $\delta^2 \leq \mu\epsilon$. The crucial part is to realize that for any iteration k such that $f(x_k) - f(x^*) > \epsilon$ (i.e. any iteration before ϵ convergence to optimum), we have by the PL Inequality that:

$$\begin{aligned} \delta^2 &\leq \mu\epsilon && (6)\\ &< \mu(f(x_k) - f(x^*)) && (7)\\ &\leq \frac{1}{2} \|\nabla f(x)\|_2^2 && (8) \end{aligned}$$

Now, at iteration k, define $\delta_k = g(x_k) - \nabla f(x_k)$. We know, $\|\delta_k\|_2 < \delta$. Assuming that $f(x_k) - f(x^*) > \epsilon$, by the Lipschitz property we have:

$$f(x_{k+1}) - f(x_k) \leq \nabla f(x_k)^\top (x_{k+1} - x_k) + \frac{L}{2} \|x_{k+1} - x_k\|_2^2 \tag{9}$$

$$= \nabla f(x_k)^\top (-\frac{1}{L} g(x_k)) + \frac{L}{2} \left\| -\frac{1}{L} g(x_k) \right\|_2^2 \tag{10}$$

$$= \nabla f(x_k)^\top (-\frac{1}{L} (\nabla f(x_k) + \delta_k)) + \frac{L}{2} \left\| -\frac{1}{L} (\nabla f(x_k) + \delta_k) \right\|_2^2 \tag{11}$$

$$= -\frac{1}{L} (\|\nabla f(x_k)\|_2^2 + \nabla f(x_k)^\top \delta_k) + \frac{1}{2L} \|\nabla f(x_k) + \delta_k\|_2^2 \tag{12}$$

$$= -\frac{1}{L} (\|\nabla f(x_k)\|_2^2 + \nabla f(x_k)^\top \delta_k) + \frac{1}{2L} (\|\nabla f(x_k)\|_2^2 + \|\delta_k\|_2^2 + 2\nabla f(x_k)^\top \delta_k) \tag{13}$$

$$= -\frac{1}{L} \|\nabla f(x_k)\|_2^2 + \frac{1}{2L} (\|\nabla f(x_k)\|_2^2 + \|\delta_k\|_2^2) \tag{14}$$

$$= -\frac{1}{2L} \|\nabla f(x_k)\|_2^2 + \frac{1}{2L} \|\delta_k\|_2^2 \tag{15}$$

$$\leq -\frac{1}{2L} \|\nabla f(x_k)\|_2^2 + \frac{1}{2L} \delta^2 \tag{16}$$

$$\leq -\frac{1}{2L} \|\nabla f(x_k)\|_2^2 + \frac{1}{2L} \frac{1}{2} \|\nabla f(x)\|_2^2 \tag{17}$$

$$= -\frac{1}{4L} \|\nabla f(x_k)\|_2^2 \tag{18}$$

Now by the PL Inequality, we have that:

$$f(x_{k+1}) - f(x_k) \leq -\frac{\mu}{2L} (f(x_k) - f(x^*)) \tag{19}$$

This then gives us:

$$f(x_{k+1}) - f(x^*) \leq (1 - \frac{\mu}{2L}) (f(x_k) - f(x^*)) \tag{20}$$

The key is that this only holds true if $f(x_k) - f(x^*) > \epsilon$. Hence if we run the algorithm for $N > \frac{2L}{\mu} \log(\frac{f(x_0) - f(x^*)}{\epsilon})$, there is no guarantee that we haven't left the ϵ ball around the optimum. A simple way to see this is to consider the case when you are sitting at the global minimum of a convex function, but are unaware of this fact. If you have a δ close estimate of the gradient, not equal to 0, your function value will get strictly worse if you take a gradient step. This is probably the really hard part of the problem. Nevertheless, the convergence measure of using the minimum that I've given here gives us an idea of what we are looking at. Below, I've given some ideas on how we could go about showing proper convergence, if we wanted to. □

If we wanted to show proper convergence and get rid of the minimum, then I have a few ideas. First, recognize that if we could simply say that we can always estimate $\nabla f(x)$ to within $\frac{1}{2} \|\nabla f(x)\|_2^2$, then we would have proper convergence. But this isn't very meaningful. So as for solutions, we essentially want to understand what argument we can make when we have $f(x_k) - f(x^*) < \epsilon$. The first thing to note is that there are really two cases to this. The first case is when $\frac{1}{2\mu} \|\nabla f(x_k)\|_2^2 > \epsilon$. Here the analysis that I have described above still holds, so no problems there. The other case is when $\frac{1}{2\mu} \|\nabla f(x_k)\|_2^2 < \epsilon$. This is the harder one to deal with.

I briefly toyed around with what would happen if we took a gradient step in this case, and got a bound on how much our function value changes, but it is not immediately obvious how this is helpful. It's definitely worth me taking a look at this again though. Another important thing to try would be to now introduce convexity, and see if that helps us in any way with this case. We could also try and define an $\epsilon' < \epsilon$. So, if we wanted to get ϵ convergence, then we run the algorithm for ϵ' convergence where ϵ' is chosen so much smaller that no matter how long we run the algorithm we never escape the ϵ ball around the optimum. Finally, note that for my result we only needed $\delta < O(\sqrt{\epsilon})$, it is worth trying $\delta < O(\epsilon)$ or $\delta < O(\epsilon^2)$.