A New Gradient Descent Optimizer (Proof)

Amish Mittal, Jimson Mathew

December 19, 2021

1 Introduction

We propose a novel gradient descent optimizer which is non-monotonic on the gradient value opposite to other descent optimizers proposed in various literature till now. The new update rule on the model parameters $\theta \in \mathbb{R}^n$ is defined as:

$$\Delta\theta_i = -\max\left(\epsilon,\ \frac{|\nabla f_i|^h}{|1+\nabla f_i^2|^h}\right).\nabla f_i \tag{1}$$

where, $\Delta\theta_i$ = parameter change along i^{th} dimension,
∇f_i = gradient of loss function along i^{th} dimension
h = hyperparameter to control convergence
ϵ = small $\mathbb{R}$ to avoid $\Delta\theta_i = 0$

2 Convergence Proofs

2.1 Convergence Validity Theorem

Theorem 2.1. *A function $f : \mathbb{R}^n -> \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$, i.e. we have $||\nabla f_x - \nabla f_y|| \leq L||x-y||_2$ for any x, y. Then, if we run gradient descent with update rule as $\Delta\theta_i = \frac{|\nabla f_i|^h}{|1+\nabla f_i^2|^h}.\nabla f_i$, it will always converge provided $h > log_2 L - 1$.*

Proof. As ∇f is Lipshitz continuos, we do a quadratic expansion of f around some value $f(x)$ to obtain:

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

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

For gradient descent step, $y = x^+ = x - \frac{|\nabla f_x|^h}{|1+\nabla f_x^2|^h}.\nabla f_x$.

$$f(x^+) \leq f(x) + \nabla f_x^T(x^+-x) + \frac{1}{2}L||x^+-x||_2^2$$

$$f(x^+) \leq f(x) + \nabla f_x^T(x - \frac{|\nabla f_x|^h}{|1+\nabla f_x^2|^h}.\nabla f_x - x) + \frac{1}{2}L||x - \frac{|\nabla f_x|^h}{|1+\nabla f_x^2|^h}.\nabla f_x - x||_2^2$$

Let $t = \frac{|\nabla f_x|^h}{|1+\nabla f_x^2|^h}.\nabla f_x$,

$$f(x^+) \leq f(x) - \nabla f(x)^T t\nabla f(x) + \frac{1}{2}L||t\nabla f(x)||_2^2$$

$$f(x^+) \leq f(x) - (1 - \frac{1}{2}Lt)\,t||\nabla f(x)||_2^2 \tag{2}$$

Now, $t||\nabla f(x)||_2^2$ will be always $+ve$ as both t and $||\nabla f(x)||_2^2$ are always $+ve$, except when $\nabla f(x)$ is 0.

For the term $(1-\frac{1}{2}Lt)$ to be $+ve$:

$$0 < 1 - \frac{1}{2}Lt$$
$$L < \frac{2}{t} \tag{3}$$

Now, $t = t(\nabla f_x)$ would be maximum at points where $\frac{d(t(\nabla f_x))}{d\nabla f_x} = 0$ and $\frac{d^2 t(\nabla f_x)}{d\nabla f_x^2} < 0$. By solving these, it can be shown that $t(\nabla f_x)$ would be maximum when $\nabla f_x = 1$ and its maximum value would be $\frac{1}{2^h}$. Hence, $\frac{2}{t}$'s minimum value would be 2^{h+1}.

$\therefore$ If $L < 2^{h+1}$ or $h > log_2 L - 1$, then $(1-\frac{1}{2}Lt)t||\nabla f(x)||_2^2$ will always be $+ve$.

From Equation 2, we can now follow that objective function value strictly decreases with each iteration of the gradient descent until it reaches the optimal value $f(x) = f(x^*)$. This result only holds if our chosen $h > log_2 L - 1$.

$$\therefore h > log_2 L - 1 \;\; or \;\; L < 2^{h+1} \;\; for \;\; convergence. \tag{4}$$

□

2.2 Convergence Rate Theorem

Theorem 2.2. *A function $f : \mathbb{R}^n -> \mathbb{R}$ is convex and differentiable, and that its gradient is Lipschitz continuous with constant $L > 0$. Then, if we run gradient descent for k iterations with update rule as $\Delta\theta_i = -\max\left(\epsilon, \frac{|\nabla f_i|^h}{|1+\nabla f_i^2|^h}\right).\nabla f_i$, it will lead to a solution $f^{(k)}$ satisfying*

$$f\left(x^{(k)}\right) - f\left(x^*\right) \leq \frac{\left\|x^{(0)} - x^*\right\|_2^2}{2\epsilon k}$$

provided that $h > log_2 L$.

Proof. We try to bound $f(x^+)$, the loss function value at the next step in terms of $f(x^*)$, the optimal value.

As f is convex:

$$f\left(x^*\right) \geq f(x) + \nabla f_x^T \left(x^* - x\right)$$
$$f(x) \leq f\left(x^*\right) + \nabla f_x^T \left(x - x^*\right)$$

Substituting this in to Equation 2, we obtain:

$$f\left(x^+\right) \leq f\left(x^*\right) + \nabla f(x)^T \left(x - x^*\right) - (1 - \frac{1}{2}Lt)t||\nabla f(x)\|_2^2$$
$$f\left(x^+\right) - f\left(x^*\right) \leq \nabla f(x)^T \left(x - x^*\right) - (1 - \frac{1}{2}Lt)t||\nabla f(x)\|_2^2$$

Taking maximum value of $L = 2^h$ and for $t = \frac{1}{2^h}$,

$$f\left(x^+\right) - f\left(x^*\right) \leq \nabla f(x)^T \left(x - x^*\right) - (1 - \frac{1}{2}2^h\frac{1}{2^h})t||\nabla f(x)\|_2^2$$
$$f\left(x^+\right) - f\left(x^*\right) \leq \nabla f(x)^T \left(x - x^*\right) - \frac{t}{2}||\nabla f(x)\|_2^2$$
$$f\left(x^+\right) - f\left(x^*\right) \leq \frac{1}{2t}\left(2t\nabla f(x)^T \left(x - x^*\right) - t^2\|\nabla f(x)\|_2^2\right)$$
$$f\left(x^+\right) - f\left(x^*\right) \leq \frac{1}{2t}\left(2t\nabla f(x)^T \left(x - x^*\right) - t^2\|\nabla f(x)\|_2^2 - \|x - x^*\|_2^2 + \|x - x^*\|_2^2\right)$$
$$f\left(x^+\right) - f\left(x^*\right) \leq \frac{1}{2t}\left(\|x - x^*\|_2^2 - \|x - t\nabla f(x) - x^*\|_2^2\right)$$

Now, for gradient descent, $x^+ = x - t\nabla f(x)$,

$$f\left(x^{+}\right)-f\left(x^{*}\right) \leq \frac{1}{2t}\left(\|x-x^{*}\|_{2}^{2}-\left\|x^{+}-x^{*}\right\|_{2}^{2}\right) \leq \frac{1}{2t_{\min}}\left(\|x-x^{*}\|_{2}^{2}-\left\|x^{+}-x^{*}\right\|_{2}^{2}\right)$$

$$f\left(x^{+}\right)-f\left(x^{*}\right) \leq \frac{1}{2\epsilon}\left(\|x-x^{*}\|_{2}^{2}-\left\|x^{+}-x^{*}\right\|_{2}^{2}\right)$$

This inequality holds for x^{+} on every epoch of gradient descent. Summing over multiple epochs, we can deduce:

$$\begin{aligned}
\sum_{i=1}^{k} f(x^{(i)})-f\left(x^{*}\right) &\leq \sum_{i=1}^{k} \frac{1}{2\epsilon}\left(\left\|x^{(i-1)}-x^{*}\right\|_{2}^{2}-\left\|x^{(i)}-x^{*}\right\|_{2}^{2}\right) \\
&= \frac{1}{2\epsilon}\left(\left\|x^{(0)}-x^{*}\right\|_{2}^{2}-\left\|x^{(k)}-x^{*}\right\|_{2}^{2}\right) \\
&\leq \frac{1}{2\epsilon}\left(\left\|x^{(0)}-x^{*}\right\|_{2}^{2}\right)
\end{aligned}$$

Using the fact that f is decreasing on every iteration, we can conclude that,

$$\begin{aligned}
f\left(x^{(k)}\right)-f\left(x^{*}\right) &\leq \frac{1}{k}\sum_{i=1}^{k} f\left(x^{(i)}\right)-f\left(x^{*}\right) \\
&\leq \frac{\left\|x^{(0)}-x^{*}\right\|_{2}^{2}}{2\epsilon k}
\end{aligned}$$

□

2.3 References

[1] https://www.stat.cmu.edu/~ryantibs/convexopt-F13/scribes/lec6.pdf