Runge Kutta Optimizers

Raghav K. Singhal

1 Introduction

In this paper we present a new optimization method, which is based on the idea that Gradient Descent is a Euler Approximation to the solution of the following Ordinary Differential Equation:

$$\dot{\theta}_t = -\nabla_\theta L(\theta_t)$$

The Euler Approximation to this Ordinary Differential Equation is of the following form:

$$x_{n+1} = x_n - \alpha \nabla_\theta L(\theta_n)$$

where α is the step-size which is the learning rate for optimization.

Look up Literature for continuous gradient descent

We explore the idea that the solution trajectory for the above Differential Equation, which also has the critical points of the loss function $L(\theta)$ as its ω-limit point.

2 RK2 - Ralston Method

Change this
for t in $[0, T]$ do **do**
  $k_1 \leftarrow \nabla f(x_t)$
  $k_2 \leftarrow \nabla f(x_t - \frac{2\alpha}{3} k_1)$
  $x_{t+1} \leftarrow x_n - \frac{\alpha}{4}(k_1 + 3k_2)$
end for

Theorem 1 (Convex Case for smooth $L(x)$ in $\mathbb{R}^d$)**.** *For some $\eta > 0$ and $k \geq 1$ and given some regularity assumptions about $f(x)$, there exists function $c(f, k)$ and $\beta \geq 0$, such that:*

$$|L(x_k) - L(x_*)| \leq ||x_k - x_*||_2^{-\eta} \frac{c(L, k)}{k^\beta}$$

To prove the above proposition we need the following lemmas:

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and satisfy $||\nabla f(x) - \nabla f(y)||$ for all $x, y \in \mathbb{R}^d$. Let $\Delta x = \frac{1}{4\beta}(k_1 + 3k_2)$ and $y = x - \Delta x$, then we show that for some $c_1 > 2\beta$*

$$f(x - \Delta x) - f(x) \leq -\frac{1}{c_1} ||\nabla f(x)||^2 \tag{1}$$

$$\tag{2}$$

Proof. Let $\Delta x = \frac{1}{4\beta}(k_1 + 3k_2)$, then

$$\begin{aligned}
f(x-\Delta x) - f(x) &\leq \nabla f(x)^T(x - \Delta x - x) + \frac{\beta}{2}||x - x - \Delta x||^2 \\
&= -\nabla f(x)^T(\Delta x) + \frac{1}{32\beta}||\Delta x||^2 \\
&= -\frac{1}{4\beta}\nabla f(x)^T(k_1 + 3k_2) + \frac{1}{32\beta}||\Delta x||^2 \\
&= -\frac{1}{4\beta}\nabla f(x)^T k_1 - \frac{3}{4\beta}\nabla f(x)^T k_2 + \frac{1}{32\beta}||k_1||_2^2 + \frac{9}{32\beta}||k_2||_2^2 + \frac{6}{32\beta}\langle k_1, k_2\rangle \\
&= -\frac{1}{4\beta}\nabla f(x)^T k_1 + \frac{1}{32\beta}||k_1||_2^2 - \frac{3}{4\beta}\nabla f(x)^T k_2 + +\frac{9}{32\beta}||k_2||_2^2 + \frac{6}{32\beta}\langle k_1, k_2\rangle \\
&= -\frac{7}{32\beta}||k_1||_2^2 - \frac{1}{32\beta}k_2^T(24\nabla f(x) - 9k_2) + \frac{6}{32\beta}\langle k_1, k_2\rangle \\
&= -\frac{7}{32\beta}||k_1||_2^2 - \frac{24}{32\beta}k_2^T k_1 + \frac{6}{32\beta}\langle k_1, k_2\rangle + \frac{9}{32\beta}||k_2||_2^2 \\
&= -\frac{7}{32\beta}||k_1||_2^2 - \frac{18}{32\beta}\langle k_1, k_2\rangle + \frac{9}{32\beta}||k_2||_2^2
\end{aligned} \tag{3}$$

Now, using a Taylor Series approximation for $\nabla f\big(x - \frac{2}{3\beta}k_1\big)$, we get that :

$$\begin{aligned}
\nabla f\big(x - \frac{2}{3\beta}k_1\big) &= \nabla f(x) - \frac{2}{3\beta}\nabla^2 f(x)\nabla f(x) + \mathcal{O}(placeholder) \\
\implies k_2^T k_1 &= \nabla f\big(x - \frac{2}{3\beta}k_1\big)^T \nabla f(x) \\
&= \nabla f\big(x - \frac{2}{3\beta}\nabla f(x)\big)^T \nabla f(x) \\
&= ||\nabla f(x)||_2^2 - \frac{2}{3\beta}\nabla f(x)^T \nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{4}$$

And, using (4),

$$\begin{aligned}
||k_2||_2^2 &= ||\nabla f(x) - \frac{2}{3\beta}\nabla^2 f(x)\nabla f(x)||_2^2 \\
&= ||\nabla f(x)||_2^2 + \frac{4}{9\beta}||\nabla^2 f(x)\nabla f(x)||_2^2 - \frac{4}{3\beta}\nabla f(x)^T\nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{5}$$

Hence, using (4) and (5)

$$
\begin{aligned}
f(x-\Delta x)-f(x) &\leq -\frac{7}{32\beta}||k_1||_2^2 - \frac{18}{32\beta}\langle k_1, k_2\rangle + \frac{9}{32\beta}||k_2||_2^2 \\
&= -\frac{7}{32\beta}||\nabla f(x)||_2^2 - \frac{18}{32\beta}||\nabla f(x)||_2^2 + \frac{12}{32\beta^2}\nabla f(x)^T\nabla^2 f(x)\nabla f(x) \\
&+ \frac{9}{32\beta}(||\nabla f(x)||_2^2 + \frac{4}{9\beta}||\nabla^2 f(x)\nabla f(x)||_2^2 - \frac{4}{3\beta}\nabla f(x)^T\nabla^2 f(x)\nabla f(x)) \\
&= -\frac{16}{32\beta}||\nabla f(x)||_2^2 + \frac{1}{8\beta^2}||\nabla^2 f(x)\nabla f(x)||_2^2 \\
&= -\frac{1}{2\beta}||\nabla f(x)||_2^2 + \frac{1}{8\beta^2}||\nabla^2 f(x)\nabla f(x)||_2^2
\end{aligned}
$$

Using the lipschitz property of the Hessian of f, $||\nabla^2 f(x)u - \nabla^2 f(x)v||_2^2 \leq \beta||u-v||_2^2$, we get that (**CHECK THE LIPSCHITZ CONSTRAINT**)

$$
\begin{aligned}
f(x-\Delta x)-f(x) &\leq -\frac{1}{2\beta}||\nabla f(x)||_2^2 + \frac{1}{8\beta^2}||\nabla^2 f(x)\nabla f(x)||_2^2 \\
&\leq -\frac{4}{8\beta}||\nabla f(x)||_2^2 + \frac{\beta}{8\beta^2}||\nabla f(x)||_2^2 \\
&= -(\frac{4}{8\beta} - \frac{\beta}{8\beta^2})||\nabla f(x)||_2^2 \\
&= -\frac{3}{8\beta}||\nabla f(x)||_2^2
\end{aligned} \tag{6}
$$

□

(Theorem 1) Proof. Using Lemma 1, we have $f(x_{t+1}) - f(x) \leq -\frac{3}{8\beta}||\nabla f(x)||_2^2$. Now, let $\delta_t = f(x_t) - f(x^*)$, then note that:

$$\delta_{t+1} \leq \delta_t - \frac{3}{8\beta}||\nabla f(x)||_2^2$$

Now, by convexity of $f(x)$ we have:

$$
\begin{aligned}
\delta_t &\leq \nabla f(x_t)^T(x_t - x^*) && (7)\\
&\leq ||x_t - x^*||_2 * ||\nabla f(x_t)||_2 && (8)\\
\frac{1}{||x_t - x^*||}\delta_t^2 &\leq ||\nabla f(x_t)||_2^2 && (9)\\
&\preceq && (10)
\end{aligned}
$$

Now, note that $||x_t - x^*||_2^2$ is decreasing, using the following

$$\big(\nabla f(x) - \nabla f(y)\big)^T(x-y) \geq \frac{1}{\beta}||\nabla f(x) - \nabla f(y)||_2^2$$

Using the above and the fact that $\nabla f(x^*) = 0$,

$$\begin{aligned}
||x_{t+1} - x^*||_2^2 &= ||x_t - \Delta x_t - x^*||_2^2 \\
&= ||x_t - x^*||_2^2 + ||\Delta x_t||_2^2 - 2\Delta x_t^T(x_t - x^*) \\
&= ||x_t - x^*||_2^2 - \frac{1}{2\beta}(k_1 + 3k_2)^T(x_t - x^*) + \frac{1}{16\beta^2}||k_1 + 3k_2||_2^2 \\
&= ||x_t - x^*||_2^2 - \frac{1}{2\beta}k_1^T(x_t - x^*) + \frac{1}{16\beta^2}||k_1||_2^2 \\
&\quad - \frac{3}{2\beta}k_2^T(x_t - x^*) + \frac{9}{16\beta^2}||k_2||_2^2 + \frac{6}{16\beta^2}k_1^T k_2 \\
&= ||x_t - x^*||_2^2 - \frac{4}{16\beta^2}||k_1||_2^2 + \frac{1}{16\beta^2}||k_1||_2^2 \\
&\quad - \frac{12}{16\beta^2}|k_2||_2^2 + \frac{9}{16\beta^2}||k_2||_2^2 + \frac{6}{16\beta^2}k_1^T k_2 \\
&= ||x_t - x^*||_2^2 - \frac{3}{16\beta^2}||k_1||_2^2 - \frac{3}{16\beta^2}||k_2||_2^2 + \frac{6}{16\beta^2}k_1^T k_2 \\
&= ||x_t - x^*||_2^2 - \frac{3}{16\beta^2}||k_1 - k_2||_2^2 \\
&\leq ||x_t - x^*||_2^2
\end{aligned}$$

We will show that,

$$\delta_{t+1} \leq \delta_t - \frac{3}{8\beta||x_1 - x^*||_2^2}\delta_t^2 \tag{11}$$

Now, let $\omega = \frac{3}{8\beta||x_1 - x^*||_2^2}$, then note that: (Proof in Bubek page - 269)

$$\begin{aligned}
&\frac{1}{\delta_t} \geq \omega(t-1) \\
\implies &f(x_t) - f(x^*) \leq \frac{8}{3\beta}\frac{||x_1 - x^*||_2^2}{t-1} \xrightarrow{t\to\infty} 0
\end{aligned}$$

□

3 Order of convergence

4 Notes

Lemma 2. *Let $f : \mathbb{R}^d \to \mathbb{R}$ be convex and satisfy $||\nabla f(x) - \nabla f(y)|| \leq \beta||x - y||$ for all $x, y \in \mathbb{R}^d$. Then $\forall x, y \in \mathbb{R}^d$ the following are true:*

$$0 \leq f(x) - f(y) - \nabla f(y)^T(x - y) \leq \frac{\beta}{2}||x - y||^2 \tag{12}$$

$$f(x) - f(y) \leq \nabla f(x)^T(x - y) - \frac{1}{2\beta}||\nabla f(x) - \nabla f(y)||^2 \tag{13}$$

$$\frac{1}{\beta}||\nabla f(x) - \nabla f(y)||^2 \leq (\nabla f(x) - \nabla f(y))^T(x - y) \tag{14}$$

4.1 Comparison to GD

$$\textbf{Remove this later} \tag{15}$$

$$f(x - \Delta x) - f(x) \leq -\frac{1}{2\beta}||\nabla f(x)||^2 \leq -\frac{1}{c_1}||\nabla f(x)||^2 \tag{16}$$

or equivalently

$$f(x) - f(x - \Delta x) \geq \frac{1}{c_1}||\nabla f(x)||^2$$

Remove this later - MAYBE SWITCH STUFF

$$f(x) - f(x - \Delta x) \geq \frac{1}{2\beta}||\nabla f(x)||^2 \geq \frac{1}{c_1}||\nabla f(x)||^2$$

Note for gradient descent $c_1 = 2\beta$, so RK2 would give a bigger step if $c_1 < 2\beta$.