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:

$$\dot{x}_t = -\nabla_x f(x_t) \tag{1}$$

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

$$x_{n+1} = x_n - \alpha \nabla_x f(x_n)$$

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 (1), which also has the critical points of the loss function $L(\theta)$ as its ω-limit point.

2 ODE Ideas - Lyupanov Function

Here we provdide certain properties of the solution of (1), when $f \in \mathcal{S}_\beta^{2,1}(\mathbb{R}^n)$, that is f is strongly convex and twice differentiable with $||\nabla f(x) - \nabla f(y)|| \leq \beta ||x - y||$, for all $x, y \in \mathbb{R}^n$. Define $r_t = f(x_t) - f(x^*)$, then note that

$$\begin{aligned} \frac{d}{dt} r_t &= \langle \nabla f(x_t), \dot{x}_t \rangle \\ &= -||\nabla f(x_t)||_2^2 \end{aligned} \tag{2}$$

Now, note that $||\nabla f(x)|| \leq \beta ||x - x^*||$, which implies that

$$\begin{aligned} -\beta^2 ||x_t - x^*|| \leq -||\nabla f(x_t)||_2^2 &= \frac{d}{dt}\big(f(x_t) - f(x^*)\big) \\ \frac{d}{dt}\big(f(x_t) - f(x^*)\big) &\geq \beta ||x_t - x^*||_2^2 \end{aligned} \tag{3}$$

But, as $f(x_t) - f(x^*)$ is continuously decreasing we have that

3 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

3.1 Proof 1

Theorem 1. *Let $f(x) \in C_\beta^{2,2}(\mathbb{R}^n) \cap C_\beta^{2,1}(\mathbb{R}^n)$ and f is bounded below, then the RK2-Ralston Method gap between x_t and some local minima x^* is given by :*

$$f(x_t) - f(x^*) \leq \frac{8}{3\beta} \frac{||x_1 - x^*||_2^2}{t-1}$$

To prove the above proposition we need the following lemmas:

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R} \in C_\beta^{2,2}(\mathbb{R}^n) \cap C_\beta^{2,1}(\mathbb{R}^n)$. Let $\Delta x = \frac{1}{4\beta}(k_1 + 3k_2)$ and $y = x - \Delta x$, then we show that, for some $c_1 = \frac{8\beta}{3} > 0$,*

$$f(x - \Delta x) - f(x) \leq -\frac{3}{8\beta} ||\nabla f(x)||^2 \tag{4}$$

$$\tag{5}$$

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

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

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

$$\begin{aligned}
\nabla f\big(x - \frac{2}{3\beta}k_1\big) &= \nabla f(x) - \frac{2}{3\beta}\nabla^2 f(x)\nabla f(x) + \mathcal{O}(|\frac{2}{3\beta}|^2) \\
\implies k_2^T k_1 &= \nabla f\big(x - \frac{2}{3\beta}k_1\big)^T \nabla f(x) \\
&= \nabla f\big(x - \frac{2}{3\beta}\nabla f(x)\big)^T \nabla f(x) \\
&= ||\nabla f(x)||_2^2 - \frac{2}{3\beta}\nabla f(x)^T \nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{7}$$

And, using (7),

$$\begin{aligned}
||k_2||_2^2 &= ||\nabla f(x) - \frac{2}{3\beta}\nabla^2 f(x)\nabla f(x)||_2^2 \\
&= ||\nabla f(x)||_2^2 + \frac{4}{9\beta}||\nabla^2 f(x)\nabla f(x)||_2^2 - \frac{4}{3\beta}\nabla f(x)^T \nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{8}$$

Hence, using (7) and (8)

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

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,

$$\begin{aligned}
f(x - \Delta x) - f(x) &\leq -\frac{1}{2\beta}||\nabla f(x)||_2^2 + \frac{1}{8\beta^2}||\nabla^2 f(x)\nabla f(x)||_2^2 \\
&\leq -\frac{4}{8\beta}||\nabla f(x)||_2^2 + \frac{\beta}{8\beta^2}||\nabla f(x)||_2^2 \\
&= -\big(\frac{4}{8\beta} - \frac{\beta}{8\beta^2}\big)||\nabla f(x)||_2^2 \\
&= -\frac{3}{8\beta}||\nabla f(x)||_2^2
\end{aligned} \tag{9}$$

□

(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:

$$\delta_t \leq \nabla f(x_t)^T(x_t - x^*) \tag{10}$$
$$\leq ||x_t - x^*||_2 * ||\nabla f(x_t)||_2 \tag{11}$$
$$\frac{1}{||x_t - x^*||}\delta_t^2 \leq ||\nabla f(x_t)||_2^2 \tag{12}$$
$$\tag{13}$$

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 $\nabla f(x^*) = 0$,

$$\begin{aligned}
||x_{t+1} - x^*||_2^2 &= ||x_t - \Delta x_t - x^*||_2^2 \\
&= ||x_t - x^*||_2^2 + ||\Delta x_t||_2^2 - 2\Delta x_t^T(x_t - x^*) \\
&= ||x_t - x^*||_2^2 - \frac{1}{2\beta}(k_1 + 3k_2)^T(x_t - x^*) + \frac{1}{16\beta^2}||k_1 + 3k_2||_2^2 \\
&= ||x_t - x^*||_2^2 - \frac{1}{2\beta}k_1^T(x_t - x^*) + \frac{1}{16\beta^2}||k_1||_2^2 \\
&\qquad - \frac{3}{2\beta}k_2^T(x_t - x^*) + \frac{9}{16\beta^2}||k_2||_2^2 + \frac{6}{16\beta^2}k_1^Tk_2 \\
&= ||x_t - x^*||_2^2 - \frac{4}{16\beta^2}||k_1||_2^2 + \frac{1}{16\beta^2}||k_1||_2^2 \\
&\qquad - \frac{12}{16\beta^2}|k_2||_2^2 + \frac{9}{16\beta^2}||k_2||_2^2 + \frac{6}{16\beta^2}k_1^Tk_2 \\
&= ||x_t - x^*||_2^2 - \frac{3}{16\beta^2}||k_1||_2^2 - \frac{3}{16\beta^2}||k_2||_2^2 + \frac{6}{16\beta^2}k_1^Tk_2 \\
&= ||x_t - x^*||_2^2 - \frac{3}{16\beta^2}||k_1 - k_2||_2^2 \\
&\leq ||x_t - x^*||_2^2
\end{aligned}$$

We will show that,

$$\delta_{t+1} \leq \delta_t - \frac{3}{8\beta||x_1 - x^*||_2^2}\delta_t^2 \tag{14}$$

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

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

$\square$

3.2 Order of Convergence

4 RK2 - Heun's Method

Heun's Method is a second order method to solving $\dot{x_t} = -\nabla f(x_t)$, and its updates are given as follows:

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

4.1 Proof 1

Let $\Delta x = \frac{1}{2\beta}(k_1 + k_2)$,

$$\begin{aligned}
f(x - \Delta x) - f(x) &\leq \nabla f(x)^T(-\Delta x) + \frac{\beta}{2}||\Delta x||_2^2 \\
&\leq -\nabla f(x)^T \Delta x + \frac{1}{8\beta}||k_1 + k_2||_2^2 \\
&\leq -\frac{1}{2\beta}||\nabla f(x)||_2^2 - \frac{1}{2\beta}\nabla f(x)^T k_2 + \frac{1}{8\beta^2}||\nabla f(x)||_2^2 + \frac{1}{8\beta}||k_2||_2^2 \\
&\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2 - \frac{1}{2\beta}\nabla f(x)^T k_2 + \frac{1}{8\beta}k_2^T k_2 \\
&\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2 - \frac{1}{8\beta}k_2^T\big(4\nabla f(x) - k_2\big) \\
&\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2 - \frac{1}{2\beta}k_2^T(\nabla f(x) - k_2) + \frac{3}{8\beta}k_2^T k_2 \\
&\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2 - \frac{1}{2\beta}k_2^T(\nabla f(x) - \nabla f(x - \frac{1}{\beta}\nabla f(x))) + \frac{3}{8\beta}||k_2||_2^2 \\
&\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2 + \frac{3}{8\beta}||k_2||_2^2 - \frac{1}{2\beta}k_2^T(\nabla f(x) - \nabla f(x - \frac{1}{\beta}\nabla f(x))
\end{aligned} \tag{15}$$

Now, note that using a Taylor Series approximation, we get that:

$$\begin{aligned}
\nabla f(x - \frac{1}{\beta}\nabla f(x)) &= \nabla f(x) - \frac{1}{\beta}\nabla^2 f(x)\nabla f(x) + \mathcal{O}(|c|^2) \\
&\implies \nabla f(x) - \nabla f(x - \frac{1}{\beta}\nabla f(x)) = \frac{1}{\beta}\nabla^2 f(x)\nabla f(x) \\
&\implies k_2^T \frac{1}{\beta}\nabla^2 f(x)\nabla f(x) = \frac{1}{\beta}\nabla f(x - \frac{1}{\beta}\nabla f(x))\nabla^2 f(x)\nabla f(x) \\
&\qquad\qquad = \frac{1}{\beta}\nabla f(x)^T\nabla f(x) - \frac{1}{\beta^2}\nabla f(x)^T\nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{16}$$

Now, combining (12) and (13), we get that:

$$
\begin{aligned}
\nabla f(x-\Delta x)-\nabla f(x) &\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2+\frac{3}{8\beta}||k_2||_2^2-\frac{1}{2\beta}k_2^T(\nabla f(x)-\nabla f(x-\frac{1}{\beta}\nabla f(x)) \\
&\leq -\frac{3}{8\beta}||\nabla f(x)||_2^2+\frac{3}{8\beta}||k_2||_2^2-\frac{1}{2\beta}||\nabla f(x)||_2^2+\frac{1}{2\beta^2}\nabla f(x)^T\nabla^2 f(x)\nabla f(x) \\
&\leq -\frac{7}{8\beta}||\nabla f(x)||_2^2+\frac{3}{8\beta}(||\nabla f(x)||_2^2+\frac{1}{\beta}||\nabla^2 f(x)\nabla f(x)||_2^2)+\frac{1}{2\beta^2}\nabla f(x)^T\nabla^2 f(x)\nabla f(\\
&\leq -\frac{1}{8\beta}||\nabla f(x)||_2^2+\frac{1}{2\beta^2}\nabla f(x)^T\nabla^2 f(x)\nabla f(x) \\
&\leq -\frac{1}{8\beta}||\nabla f(x)||_2^2+\frac{1}{2\beta^2}
\end{aligned}
\tag{17}
$$

4.2 Order Of Convergence

Proof. Let $r_k = ||x_k - x^*||$, the note that

$$
\begin{aligned}
x_{k+1}-x^* &= x_k - x^* - \frac{1}{2\beta}\big(\nabla f(x_k)+\nabla f(x_k-\frac{1}{\beta}\nabla f(x_k))\big) \\
&= x_k - x^* - \frac{1}{2\beta}\big(\nabla f(x_k)-\nabla f(x^*)\big)-\frac{1}{2\beta}\big(\nabla f(x_k-\frac{1}{\beta}\nabla f(x_k))-\nabla f(x^*)\big) \\
&= x_k - x^* - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+t(x_k-x^*)(x_k-x^*)dt \\
&\qquad -\frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+t(x_k-x^*-\frac{1}{\beta}\nabla f(x_k))(x_k-x^*-\frac{1}{\beta}\nabla f(x_k))dt
\end{aligned}
\tag{18}
$$

Now, let $z_k = x_k - x^*$, then note that

$$
\begin{aligned}
y_{k+1} &= y_k - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+ty_k)y_k dt - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))(y_k-\frac{1}{\beta}\nabla f(x_k))dt \\
&= \big(I-\frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+ty_k)+\nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))dt\big)y_k \\
&\qquad +\frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))\frac{1}{\beta}\nabla f(x_k)dt \\
&= \big(I-\frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+ty_k)+\nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))dt\big)y_k \\
&\qquad +\frac{1}{2\beta^2}\int_0^1 \nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))(\nabla f(x_k)-\nabla f(x^*))dt
\end{aligned}
\tag{19}
$$

Now, define the following operators:

$$
\begin{aligned}
H_k &= \big(I-\frac{1}{2\beta}\int_0^1 \nabla^2 f(x^*+ty_k)+\nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))dt\big) \\
G_k &= \frac{1}{2\beta^2}\int_0^1 \nabla^2 f(x^*+t(y_k-\frac{1}{\beta}\nabla f(x_k))dt
\end{aligned}
\tag{20}
$$

Then note that as $f \in C_{\beta}^{2,2}(\mathbb{R}^n) \cap C_{\beta}^{2,1}(\mathbb{R}^n)$,

$$||\nabla^2 f(x)|| \leq \beta \tag{21}$$
$$||\nabla f(x)|| = ||\nabla f(x) - \nabla f(x^*)|| \leq \beta ||x - x^*|| \tag{22}$$

Now, using (21) and (22), we have:

$$\begin{aligned} ||G_k(\nabla f(x_k) - \nabla f(x^*))|| &\leq ||G_k|| * ||\nabla f(x_k) - \nabla f(x^*)|| \\ &\leq \frac{1}{2\beta^2}\int_0^1 ||\nabla^2 f(x^* + t(y_k - \frac{1}{\beta}\nabla f(x_k))|| * ||\nabla f(x_k) - \nabla f(x^*)||dt \\ &\leq \frac{1}{2\beta^2}\int_0^1 \beta^2 ||x_k - x^*||dt = \frac{1}{2}||x_k - x^*|| = \frac{1}{2}r_k \end{aligned} \tag{23}$$

Note that if $||x - y|| = r$, then for $f \in C_{\beta}^{2,2}(\mathbb{R}^n)$,

$$\nabla^2 f(x) - \beta r I \preceq \nabla^2 f(y) \preceq \nabla^2 f(x) + \beta r I \tag{24}$$

And, a similar inequality can be derived for H_k, and assuming that $lI \preceq \nabla^2 f(x^*)$

$$\begin{aligned} H_k &= I - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^* + ty_k) + \nabla^2 f(x^* + t(y_k - \frac{1}{\beta}\nabla f(x_k))dt \\ &\leq I + \frac{1}{\beta}\nabla^2 f(x^*) - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^* + ty_k) - \nabla^2 f(x^*) + \nabla^2 f(x^* + t(y_k - \frac{1}{\beta}\nabla f(x_k)) - \nabla^2 f(x^*)dt \\ ||H_k|| &\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{1}{2\beta}\int_0^1 ||ty_k||dt + \frac{1}{2\beta}\int_0^1 ||ty_k - \frac{t}{\beta}\nabla f(x_k))||dt \\ &\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{4\beta} + \frac{1}{2\beta}\int_0^1 ||ty_k - \frac{t}{\beta}\nabla f(x_k))||dt \\ &\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{4\beta} + \frac{1}{2\beta}\int_0^1 t||y_k - \frac{1}{\beta}\nabla f(x_k))||dt \\ &\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{4\beta} + \frac{1}{2\beta}\int_0^1 ||ty_k||dt \\ &\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{2\beta} \\ &\leq \frac{1}{\beta}\lambda_{\max}(\nabla^2 f) + 1 + \frac{r_k}{4\beta} \end{aligned} \tag{25}$$

Using, the above inequalities we get:

$$\begin{aligned} y_{k+1} &= H_k y_k + G_k \\ r_{k+1} &\leq ||H_k||r_k + ||G_k|| \\ &\leq (\frac{1}{\beta}\lambda_{\max}(\nabla^2 f) + \frac{3}{2})r_k + \frac{r_k^2}{4\beta} \\ r_{k+1} &\leq \mu_1 r_k + \mu_2 r_k^2 \end{aligned} \tag{26}$$

□

5 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:*

$$0 \leq f(x) - f(y) - \nabla f(y)^T (x - y) \leq \frac{\beta}{2} ||x - y||^2 \tag{27}$$

$$f(x) - f(y) \leq \nabla f(x)^T (x - y) - \frac{1}{2\beta} ||\nabla f(x) - \nabla f(y)||^2 \tag{28}$$

$$\frac{1}{\beta} ||\nabla f(x) - \nabla f(y)||^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y) \tag{29}$$

5.1 Comparison to GD