Gradient Flow 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 and in the Optimization Literature is referred to as the Learning Rate. In recent years there has been some interest in analyzing Optimzation Schemes using Differential Equation [1], etc.

Note [1] uses a second-order differential equation to analyze Nesterov's Accelerated Gradient Descent, but it reverse-engineers a Differential Equation to analyze a particular method. We take the forward approach of first motivating the Ordinary Differential Equation point of view and then deriving optimization schemes, an approach already taken by [2], where they analyze linear-multistep methods and other methods for integration of Differential Equations and analyze their performance as Optimization Schemes. We look at a very famous and stable class of optimizers, Runge-Kutta Methods, and compare their performance with Gradient Descent, AdaGrad, Stochastic Gradient Descent, SGD with momentum and Accelerated Stochastic Gradient Descent.

However, the goal in each field is different, in Optimization we look at an infinite-time horizon to find an approximation close to a global or a local minima but in Numerical Analysis of Differential Equations, the goal is to find a close approximation to the solution of the Differential Equation over a finite time interval $[0, T]$. But we try to bridge this gap by saying that the the solution trajectory for (1) stays close if not converges to the critical points of f, and we formalize this view by noting that the critical points of the loss function $f(x)$ are also the ω-limit points of the (1).

2 Motivation

Here we provdide certain properties of the solution of (1), when $f \in \mathcal{S}^{2,1}_{\mu,\beta}(\mathbb{R}^n)$, that is f is strongly convex and twice differentiable with $||\nabla f(x) - \nabla f(y)|| \leq \beta ||x-y||$, for all $x, y \in \mathbb{R}^n$ and f is μ-strongly convex. Then note that $\forall t > 0$:

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

But, as $f(x_t) \in \mathcal{S}^{2,1}_{\mu,\beta}(\mathbb{R}^n)$, we have that:

$$f(x) - f(x^*) \leq \frac{1}{2\mu}||\nabla f(x)||_2^2 \tag{3}$$

Hence,

$$\begin{aligned} \frac{d}{dt}\big(f(x_t) - f(x^*)\big) &\leq -2\mu\big(f(x_t) - f(x^*)\big) \\ \implies \big(f(x_t) - f(x^*)\big) &\leq e^{-2\mu t}(f(x_0) - f(x^*)) \end{aligned} \tag{4}$$

Hence, $\forall \epsilon > 0$, there exist $t > 0$, such that $f(x_t) - f(x^*) \leq \epsilon$. And note that if we start with 2 different initial conditions, x_0^1 and x_0^2, they too converge to the same value. More precisely, let $\mathcal{L}(t) = \|x_1(t) - x_2(t)\|^2$, then note that:

$$\begin{aligned} \frac{d}{dt}\mathcal{L}(t) &= 2(x_1(t) - x_2(t))^T(\dot{x_1}(t) - \dot{x_2}(t)) \\ &= 2(x_1(t) - x_2(t))^T(-\nabla f(x_1(t)) + \nabla f(x_2(t))) \\ &= -2(x_1(t) - x_2(t))^T(\nabla f(x_1(t)) - \nabla f(x_2(t))) \\ &\leq -2\mu\|x_1(t) - x_2(t)\|^2 \\ &= -2\mu\mathcal{L}(t) \end{aligned} \tag{5}$$

As we show below, $\mathcal{L}(t)$ is decreasing so we have that:

$$\frac{d}{dt}\mathcal{L}(t) \leq -2\mu\mathcal{L}(0) \tag{6}$$

which implies that $\mathcal{L}(t) \leq e^{-2\mu t}\|x_1(0) - x_2(0)\|^2$. Giving us the following result that

Proposition 1. *Let $f \in \mathcal{S}^{2,1}_{\mu,\beta}$. Let x^* be the global minimum of f, then the solution of $\dot{x}_t = -\nabla f(x_t)$ satisfies:*

$$\begin{aligned} f(x_t) - f(x^*) &\leq e^{-2\mu t}\big(f(x_0) - f(x^*)\big) \\ \|x_t - x^*\|^2 &\leq e^{-2\mu t}\|x_0 - x^*\|^2 \end{aligned}$$

This leads to the question, whether higher order integration methods for Ordinary Differential Equation might lead to better optimization schemes. The Euler scheme for integration of O.D.E.'s can be derived by a simple Taylor Series formula,

$$x_{t+\Delta t} = x_t - \Delta t \nabla f(x_t) + \mathcal{O}(\Delta t^2)$$

There are some other phenomenon that we can study and connect with Optimization, mainly stability of Solutions and Stiffness, which we do not delve into in this report. For reference to these ideas, please look at [2].

3 Methods

In this report, the results are in a non-convex setting where $f \in \mathcal{C}_{\beta}^{1,1} \cap \mathcal{C}_{\beta}^{2,2}$, that is f is twice-differentiable and has the same lipshcitz constant β for both the first and second derivative, as well as being bounded below. (Note we do not assume convexity)

In the proceeding sections, we present algorithms and convergence proofs for two well known methods of integration, RK2-ralston method and RK2-Heun's Method,which belong to the Runge-Kutta Method Familiy.

In the experiments we did for this project, we implement RK2 Heun and RK2 Ralston along with the classical and most well known method of the Runge-Kutta Family, which is commonly called RK4 as it is a fourth order method of integrating Ordinary Differential Equations.

4 RK2 - Ralston Method

Here we present the first Runge-Kutta Method, a $2nd$ order method also known as RK2-Ralston, which we refer to as RK2 in the experiments.

```
Given $x_0$
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
```

4.1 Main Results

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, where $\alpha = \frac{2}{\beta}$:*

$$f(x_t) - f(x^*) \leq \frac{4}{3\beta} \frac{||x_1 - x^*||_2^2}{t-1}$$

To prove the above proposition we need the following lemma, where we show the amount of progress made by our integration scheme in 1 step.

Lemma 1. *Let $f : \mathbb{R}^d \to \mathbb{R} \in C_{\beta}^{2,2}(\mathbb{R}^n) \cap C_{\beta}^{2,1}(\mathbb{R}^n)$ and f be bounded below. Let $\Delta x = \frac{1}{2\beta}(k_1 + 3k_2)$, then we show that,*

$$f(x - \Delta x) - f(x) \leq -\frac{3}{4\beta}||\nabla f(x)||^2 \tag{7}$$

Proof. (Lemma 1) Let $\Delta x = \frac{1}{2\beta}(k_1 + 3k_2)$, where $\alpha = \frac{2}{\beta}$ then

$$\begin{aligned}
f(x-\Delta x) - f(x) &\leq \nabla f(x)^T(x - \Delta x - x) + \frac{\beta}{2}||x - x - \Delta x||^2 \\
&= -\nabla f(x)^T(\Delta x) + \frac{1}{16\beta}||\Delta x||^2 \\
&= -\frac{1}{2\beta}\nabla f(x)^T(k_1 + 3k_2) + \frac{1}{16\beta}||\Delta x||^2 \\
&= -\frac{1}{2\beta}\nabla f(x)^T k_1 - \frac{3}{2\beta}\nabla f(x)^T k_2 + \frac{1}{16\beta}||k_1||_2^2 + \frac{9}{16\beta}||k_2||_2^2 + \frac{6}{16\beta}\langle k_1, k_2\rangle \\
&= -\frac{1}{2\beta}\nabla f(x)^T k_1 + \frac{1}{16\beta}||k_1||_2^2 - \frac{3}{2\beta}\nabla f(x)^T k_2 + \frac{9}{16\beta}||k_2||_2^2 + \frac{6}{16\beta}\langle k_1, k_2\rangle \\
&= -\frac{7}{16\beta}||k_1||_2^2 - \frac{1}{16\beta}k_2^T(24\nabla f(x) - 9k_2) + \frac{6}{16\beta}\langle k_1, k_2\rangle \\
&= -\frac{7}{16\beta}||k_1||_2^2 - \frac{24}{16\beta}k_2^T k_1 + \frac{6}{16\beta}\langle k_1, k_2\rangle + \frac{9}{16\beta}||k_2||_2^2 \\
&= -\frac{7}{16\beta}||k_1||_2^2 - \frac{18}{16\beta}\langle k_1, k_2\rangle + \frac{9}{16\beta}||k_2||_2^2
\end{aligned} \tag{8}$$

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

$$\begin{aligned}
\nabla f\big(x - \frac{2\alpha}{3}k_1\big) &= \nabla f(x) - \frac{2\alpha}{3}\nabla^2 f(x)\nabla f(x) + \mathcal{O}(|\frac{2\alpha}{3}|^2) \\
\implies k_2^T k_1 &= \nabla f\big(x - \frac{2\alpha}{3}k_1\big)^T \nabla f(x) \\
&= \nabla f\big(x - \frac{2\alpha}{3}\nabla f(x)\big)^T \nabla f(x) \\
&= ||\nabla f(x)||_2^2 - \frac{2\alpha}{3}\nabla f(x)^T \nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{9}$$

And, using (9),

$$\begin{aligned}
||k_2||_2^2 &= ||\nabla f(x) - \frac{2\alpha}{3}\nabla^2 f(x)\nabla f(x)||_2^2 \\
&= ||\nabla f(x)||_2^2 + \frac{4\alpha}{9}||\nabla^2 f(x)\nabla f(x)||_2^2 - \frac{4\alpha}{3}\nabla f(x)^T\nabla^2 f(x)\nabla f(x)
\end{aligned} \tag{10}$$

Hence, using (9) and (10)

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

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,

$$\begin{aligned}
f(x - \Delta x) - f(x) &\leq -\frac{1}{\beta}||\nabla f(x)||_2^2 + \frac{1}{4\beta^2}||\nabla^2 f(x)\nabla f(x)||_2^2 \\
&\leq -\frac{4}{4\beta}||\nabla f(x)||_2^2 + \frac{\beta}{4\beta^2}||\nabla f(x)||_2^2 \\
&= -\big(\frac{4}{4\beta} - \frac{\beta}{8\beta^2}\big)||\nabla f(x)||_2^2 \\
&= -\frac{3}{4\beta}||\nabla f(x)||_2^2
\end{aligned} \tag{11}$$

□

Proof. (Theorem 1) Using Lemma 1, we have $f(x_{t+1}) - f(x) \leq -\frac{3}{4\beta}||\nabla f(x)||_2^2$. Now, let $\delta_t = f(x_t) - f(x^*)$, then note that:

$$\delta_{t+1} \leq \delta_t - \frac{3}{4\beta}||\nabla f(x)||_2^2$$

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

$$\begin{aligned}
\delta_t &\leq \nabla f(x_t)^T(x_t - x^*) &(12)\\
&\leq ||x_t - x^*||_2 * ||\nabla f(x_t)||_2 &(13)\\
\frac{1}{||x_t - x^*||}\delta_t^2 &\leq ||\nabla f(x_t)||_2^2 &(14)
\end{aligned}$$

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

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

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

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

We will show that,

$$\delta_{t+1} \leq \delta_t - \frac{3}{4\beta||x_1 - x^*||_2^2}\delta_t^2 \tag{15}$$

Now, let $\omega = \frac{3}{4\beta||x_1 - x^*||_2^2}$,

$$\begin{aligned}
&\frac{1}{\delta_t} \geq \omega(t-1) \\
\implies &f(x_t) - f(x^*) \leq \frac{4}{3\beta}\frac{||x_1 - x^*||_2^2}{t-1} \xrightarrow{t\to\infty} 0
\end{aligned}$$

□

5 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:

```
Given $x_0$
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
```

5.1 Main Results

Theorem 2. *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 :*

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

To prove the above proposition we need the following lemma, where we show the amount of progress made by our integration scheme in 1 step.

Lemma 2. *Let $f : \mathbb{R}^d \to \mathbb{R} \in C_{\beta}^{2,2}(\mathbb{R}^n) \cap C_{\beta}^{2,1}(\mathbb{R}^n)$ and f be bounded below. Let $\Delta x = \frac{1}{\beta}(k_1 + k_2)$, then we show that,*

$$f(x - \Delta x) - f(x) \leq -\frac{3}{2\beta} ||\nabla f(x)||^2 \tag{16}$$

We provide the proof in the appendix alongwith a proof showing that RK2 methods achieve a linear rate of convergence.

6 Experiments

6.1 Deep Learning

In all our experiments with deep learning models, we saw that for Runge-Kutta methods the training loss decereases extremely fast, as you can see it decreases to near zero in less than 2 epochs, this phenomenon was observed across all the models we tried but we can not show why we observe and under what conditions are we going to observe.

Another interesting observation we say across all the models we deployed is that the training loss is not too sensitive to changes in the learning rate, in other words for most learning rates the train loss goes to zero and extremely fast.

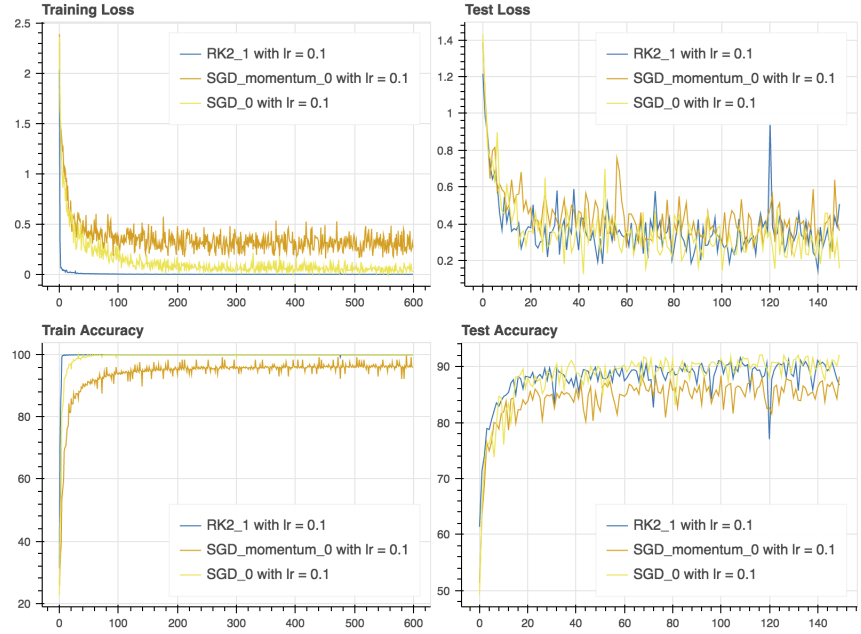
We also note that RK2 is less noisy compared to Stochastic Gradient Descent, that is the fluctuations in the loss for RK2 have less variance than the fluctuations for the loss in Stochastic Gradient Descent.

However the bad news so far is that despite training loss decreasing extremely fast, it tends to overfit. Here, we observe that if we decay the learning rate fast enough then there is a substantial increase in test performance.

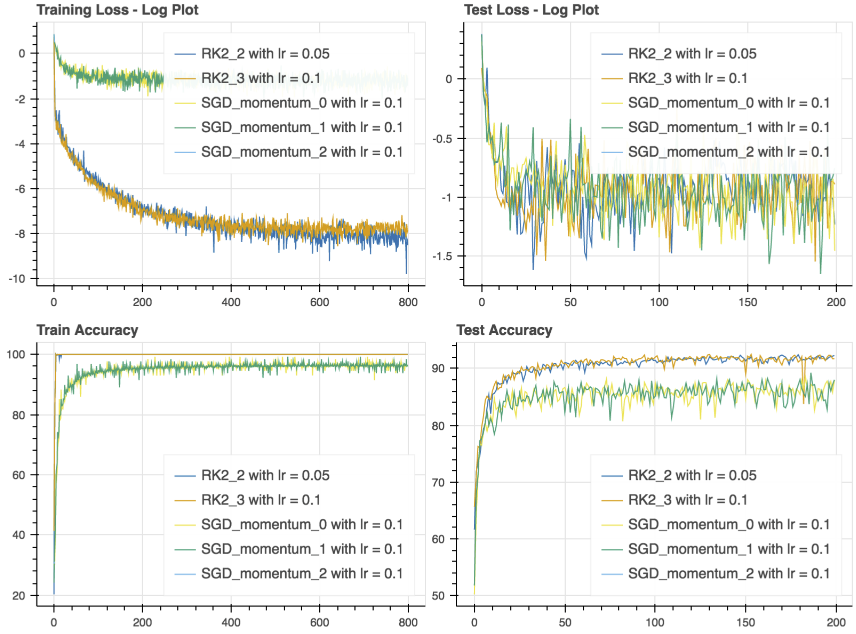
Here, we present experimental results, where without any weight-decay Runge-Kutta Method constantly outperform Stochastic Gradient Descent, with momentum or nesterov.

Model	RK2-Test Accuracy	RK2-Test Loss	SGD-Test Accuracy	SGD-Test Loss
WideResNet	93.07%	0.286	92.95%	0.249
ResNet18	92.5%	0.212	89.27%	0.1914
Logistic Regression	92.14%	0.0015	91.95%	0.00101
Lasso	92.65%	0.0021	92.52%	0.00210

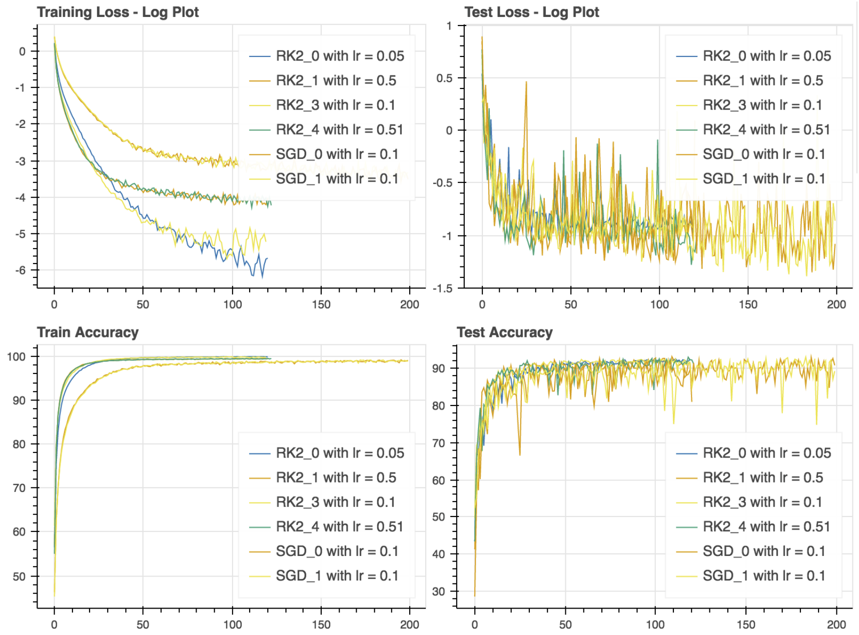
ResNet18 on Cifar10 - Here we compare our optimization scheme with Stochastic Gradient Descent with momentum and learning rate decay using the Resnet18 network [4].
Now, we compare RK2-Ralston with plain SGD and SGD with momentum



Now, we compare RK2-Heun with plain SGD and SGD with momentum



WideResnet on Cifar10 - Here we compare our optimization scheme, Runge-Kutta Ralston Method with Stochastic Gradient Descent, on the model WideResnet [5]. The different Runge-Kutta plots correspond to different learning rate decay schedules and different learning rates. Here again we can say that the variance on the test loss and test accuracy is much less compared to SGD with momentum. And also RK2 is not too sensitive to learning rate changes.

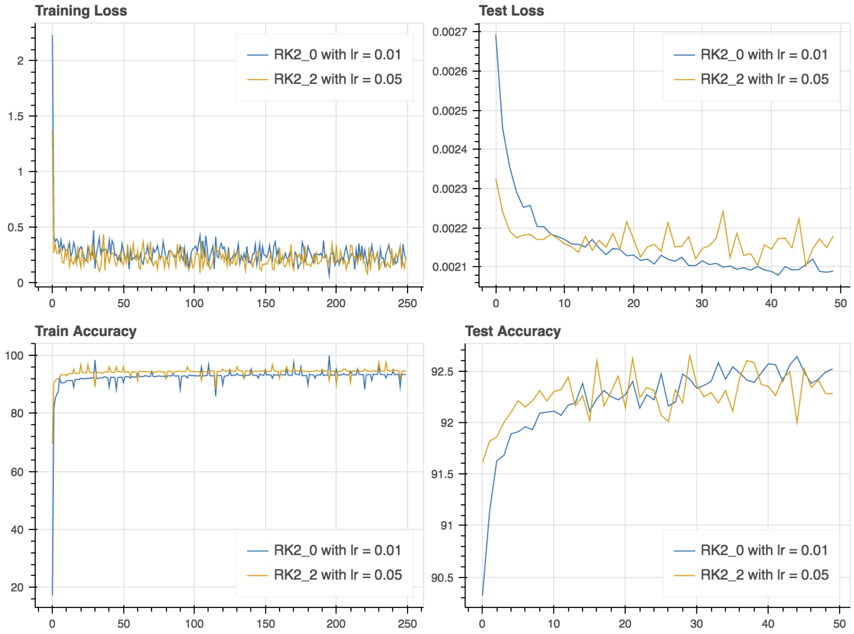


6.2 Convex Models

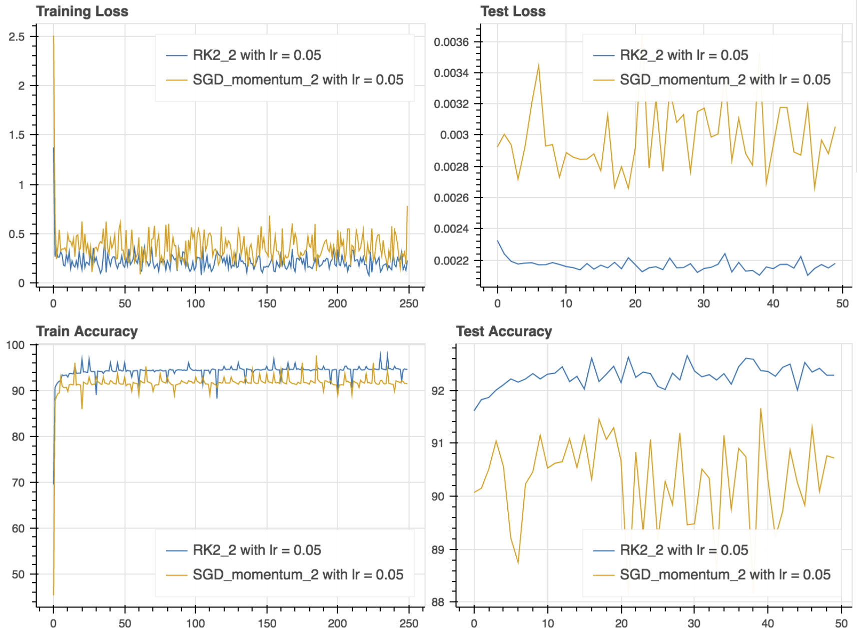
Here we do plain logistic regression and logistic regression with weight-decay on the mnist dataset. A big difference between the the previous experiments and these experiments is that the number of parameters in the model here are less than the number of training samples. So the chances of overfitting are much less. Here we present some plots for the experiments we did with these models.

Lasso

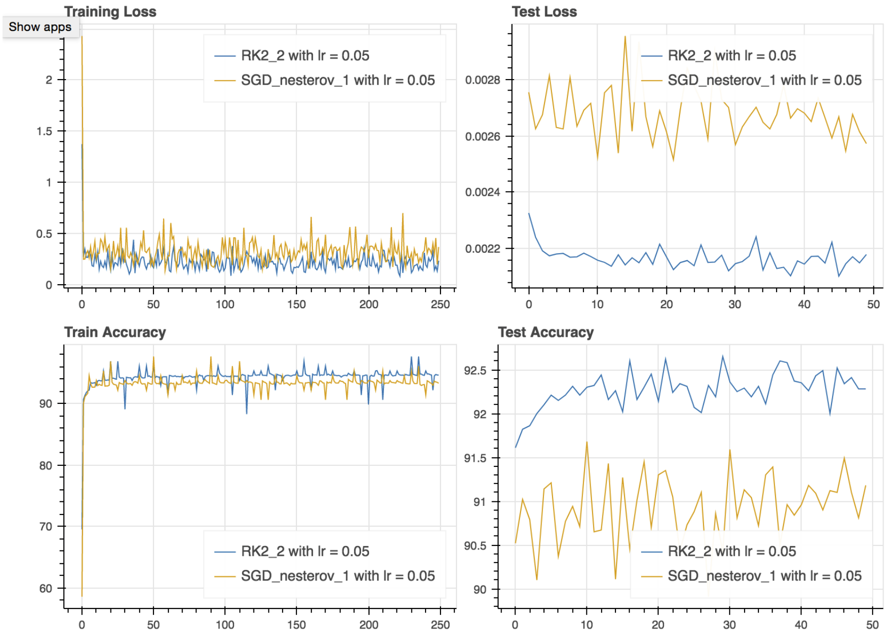
Here we show that our algorithm is not too sensitive to changes in the learning rate



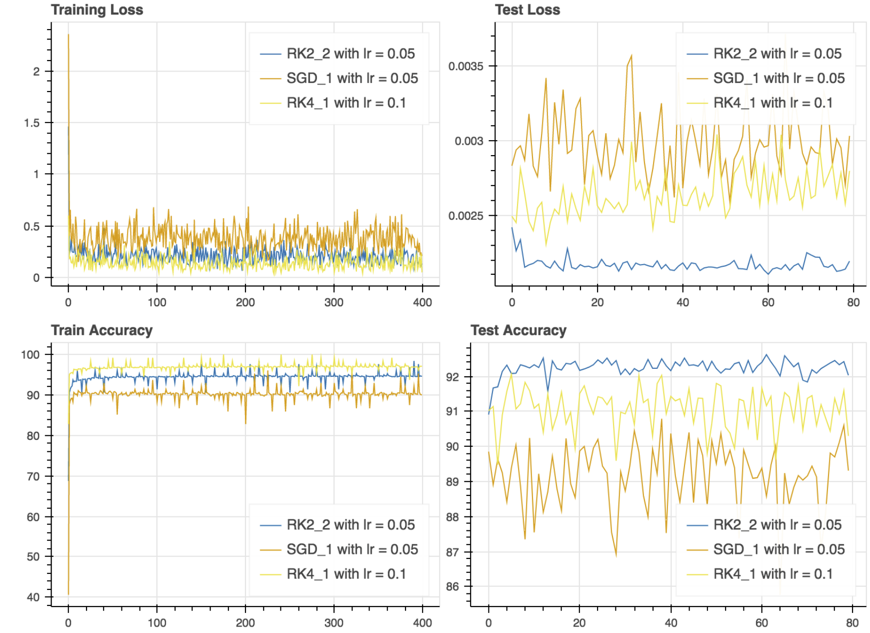
Now, we show plots to compare our optimizer with simple stochastic gradient descent, its variants and Adagrad.



Below we show plots to compare our optimizer with Accelerated stochastic gradient descent.



Here we compare a 4th order optimizer based on Runge-Kutta Methods that we wrote with RK2 and SGD



7 Conclusion and Further Work

In this work we have provided theoretical proofs and experimental justification for these methods to be studied further. Our proofs do not exactly explain the behaviour we observe in the experiments, primarily how does the training loss decrease much faster than Stochastic Gradient Descent.

One of the most interesting questions that has arisen from this study for us is under what circumstances should we choose a particular optimizer. Optimization in Deep Learning is a double-edged sword as not only is non-convex but with certain activation functions, it is not even smooth. Hence, we propose to understand Optimization in Non-Convex settings through a similar analysis applied to the Integration Schemes for Ordinary Differential Equations.

Another interesting consequence of this study and of [2], is to see if optimization techniques under convex settings can lead to faster solver for Ordinary Differential Equations.

Appendix

7.1 RK2 Heun Proof

Proof. (Lemma 2) Let $\Delta x = \frac{1}{\beta}(k_1 + k_2)$, then using Taylor Series approximation we get that:

$$\begin{aligned} f(x - \Delta x) - f(x) &\leq \nabla f(x)^T(-\Delta x) + \frac{\beta}{2}||\Delta x||_2^2 \\ &= -\frac{1}{\beta}\nabla f(x)^T(k_1 + k_2) + \frac{1}{2\beta}||k_1 + k_2||_2^2 \\ &= -\frac{1}{\beta}\nabla||f(x)||_2^2 - \frac{1}{\beta}\nabla f(x)^T k_2 + \frac{1}{2\beta}||\nabla f(x)||_2^2 + \frac{1}{2\beta}||k_2||_2^2 + \frac{1}{\beta}k_1^T k_2 \\ &\leq -\frac{1}{2\beta}||\nabla f(x)||_2^2 + \frac{1}{\beta}||k_2||_2^2 \\ &= -\frac{1}{2\beta}||\nabla f(x)||_2^2 + \frac{1}{\beta}||\nabla f(x)||_2^2 + \frac{4}{\beta^2}||\nabla^2 f(x)\nabla f(x)||_2^2 \\ &\leq -\frac{3}{2\beta}||\nabla f(x)||_2^2 \end{aligned} \tag{17}$$

□

7.2 Order Of Convergence

Here we show that RK2-Heun and also RK2-ralston both achieve a linear rate of convergence.

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

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

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

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

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:

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

And now we obtain a similar bound for H_k,

$$\begin{aligned}
H_k &= I - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^* + ty_k) + \nabla^2 f(x^* + t(y_k - \frac{1}{\beta}\nabla f(x_k))dt \\
&\leq I + \frac{1}{\beta}\nabla^2 f(x^*) - \frac{1}{2\beta}\int_0^1 \nabla^2 f(x^* + ty_k) - \nabla^2 f(x^*) + \nabla^2 f(x^* + t(y_k - \frac{1}{\beta}\nabla f(x_k)) - \nabla^2 f(x^*)dt \\
||H_k|| &\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{1}{2\beta}\int_0^1 ||ty_k||dt + \frac{1}{2\beta}\int_0^1 ||ty_k - \frac{t}{\beta}\nabla f(x_k))||dt \\
&\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{4\beta} + \frac{1}{2\beta}\int_0^1 ||ty_k - \frac{t}{\beta}\nabla f(x_k))||dt \\
&\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{4\beta} + \frac{1}{2\beta}\int_0^1 t||y_k - \frac{1}{\beta}\nabla f(x_k))||dt \\
&\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{4\beta} + \frac{1}{2\beta}\int_0^1 ||ty_k||dt \\
&\leq ||I + \frac{1}{\beta}\nabla^2 f(x^*)|| + \frac{r_k}{2\beta} \\
&\leq \frac{1}{\beta}\lambda_{\max}(\nabla^2 f) + 1 + \frac{r_k}{4\beta}
\end{aligned} \tag{24}$$

Using, the above inequalities we get:

$$\begin{aligned}
y_{k+1} &= H_k y_k + G_k \\
r_{k+1} &\leq ||H_k|| r_k + ||G_k|| \\
&\leq (\frac{1}{\beta}\lambda_{\max}(\nabla^2 f) + \frac{3}{2}) r_k + \frac{r_k^2}{4\beta} \\
r_{k+1} &\leq \mu_1 r_k + \mu_2 r_k^2
\end{aligned} \tag{25}$$

7.3 Inequalities

Here we present some common inequalities in convex analysis without proof, for proof refer to [3]

Lemma 3. *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:*

$$f(x) - f(y) \leq \nabla f(x)^T (x - y) - \frac{1}{2\beta}||\nabla f(x) - \nabla f(y)||^2 \tag{26}$$

$$\frac{1}{\beta}||\nabla f(x) - \nabla f(y)||^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y) \tag{27}$$

Now, for the class of Strongly Convex functions $\mathcal{S}^{k,p}_{\beta,\mu}(\mathbb{R}^n)$ we use the following inequalities:

$$
\begin{aligned}
\langle \nabla f(x) - \nabla f(y), x - y \rangle \geq 0 \\
0 \leq f(y) - f(x) - \nabla f(x)^T (y - x) \leq \frac{\beta}{2} ||x - y||_2^2 \\
f(x) + \nabla f(x)^T (y - x) + \frac{1}{2\beta} ||\nabla f(x) - \nabla f(y)||_2^2 \leq f(y) \\
\frac{1}{\beta} ||\nabla f(x) - \nabla f(y)||_2^2 \leq (\nabla f(x) - \nabla f(y))^T (x - y) \leq \beta ||x - y||_2^2 \\
f(y) \geq f(x) + \nabla f(x)^T (y - x) + \frac{1}{2\mu} ||y - x||_2^2
\end{aligned}
\tag{28}
$$

Note, that $f(x) \geq f(x^*) + \frac{1}{2\mu} ||x - x^*||_2^2$

References

[1] Su, Weijie, Stephen Boyd, and Emmanuel Candes. *A differential equation for modeling Nesterov's accelerated gradient method: Theory and insights*. Advances in Neural Information Processing Systems. 2014.

[2] Scieur, Damien, et al. *Integration Methods and Accelerated Optimization Algorithms* arXiv preprint arXiv:1702.06751 (2017).

[3] Nesterov, Yurii. *Introductory lectures on convex optimization: A basic course*. Vol. 87. Springer Science and Business Media, 2013.

[4] Resnet18,`https://github.com/kuangliu/pytorch-cifar`

[5] WideResnet,`https://github.com/szagoruyko/wide-residual-networks/tree/master/pytorch`