Mathematics 2:
Homework 2

David Ocepek

March 31, 2022

1 Periodicity of GD variants

All three GD variants can exhibit periodic behaviour. For GD we show analytically that it can loop given any strictly convex function f. For PolyakGD and NestorovGD we find a set of parameters on $f(x) = x^2$ where the gradients loop.

1.1 GD

$$
\begin{aligned}
&x_1 = x_0 - \gamma \nabla f(x_0) \\
&x_2 = x_1 - \gamma \nabla f(x_1) \\
&x_0 = x_2 = x_0 - \gamma \nabla f(x_0) - \gamma \nabla f(x_0 - \gamma \nabla (x_0)) \\
&\quad \nabla f(x_0) = -\nabla f(x_0 - \gamma \nabla (x_0)) \\
&\quad 2x_0 = \gamma \nabla (x_0) + C \\
&\quad \nabla (x_0) = \frac{2x_0 - C}{\gamma} \\
&\quad \textit{Example} : f(x) = x^2, \gamma = 1, x_0 \in R
\end{aligned}
\tag{1}
$$

1.2 PolyakGD

$$
\begin{aligned}
&x_2 = x_1 - \gamma \nabla f(x_1) + \mu(x_1 - x_0) \\
&x_0 = x_2 - \gamma \nabla f(x_2) + \mu(x_2 - x_1) \\
&x_1 = x_0 - \gamma \nabla f(x_0) + \mu(x_0 - x_2) \\
&\quad \textit{Solution} : \gamma = 1.5, \mu = 1, x_0 = 0, x_2 = -x_1
\end{aligned}
\tag{2}
$$

1.3 NesterovGD

$$
\begin{aligned}
&x_2 = x_1 - \gamma \nabla f(x_1 + \mu(x_1 - x_0)) + \mu(x_1 - x_0) \\
&x_0 = x_2 - \gamma \nabla f(x_2 + \mu(x_2 - x_1)) + \mu(x_2 - x_1) \\
&x_1 = x_0 - \gamma \nabla f(x_0 + \mu(x_0 - x_2)) + \mu(x_0 - x_2) \\
&\quad \textit{Solution} : \gamma \approx 1.5, \mu \approx -0.5, x_0 = 0, x_2 = -x_1
\end{aligned}
\tag{3}
$$

2 Optimal Learning Rate for Polyak GD

$$
\begin{aligned}
f(x,y,z) &= x^2 + 2y^2 - 2yz + 4z^2 + 3x - 4y + 5z \\
\nabla f(x,y,z) &= \begin{bmatrix} 2x+3, & 4y-2z-4, & 8z-2y+5 \end{bmatrix}^T \\
\nabla^2 f(x,y,z) &= \begin{bmatrix} 2 & 0 & 0 \\ 0 & 4 & -2 \\ 0 & -2 & 8 \end{bmatrix} \\
det(\nabla^2 f - \lambda I) &= (2-\lambda) \begin{vmatrix} 4-\lambda & -2 \\ -2 & 8-\lambda \end{vmatrix} \\
\begin{vmatrix} 4-\lambda & -2 \\ -2 & 8-\lambda \end{vmatrix} &= \lambda^2 - 12\lambda + 28 \qquad (4) \\
\lambda_{2,3} &= \frac{12 \pm \sqrt{144 - 4*28}}{2} = 6 \pm 2\sqrt{2} \\
\lambda_{min} = \alpha = 2, \lambda_{max} &= \beta = 6 + 2\sqrt{2}, \\
\gamma &= \frac{4}{\left(\sqrt{\alpha} + \sqrt{\beta}\right)^2} = 0.2079819 \\
\mu &= \left(\frac{\sqrt{\beta} - \sqrt{\alpha}}{\sqrt{\beta} + \sqrt{\alpha}}\right)^2 = 0.1260586
\end{aligned}
$$

3 Comparison of GD methods

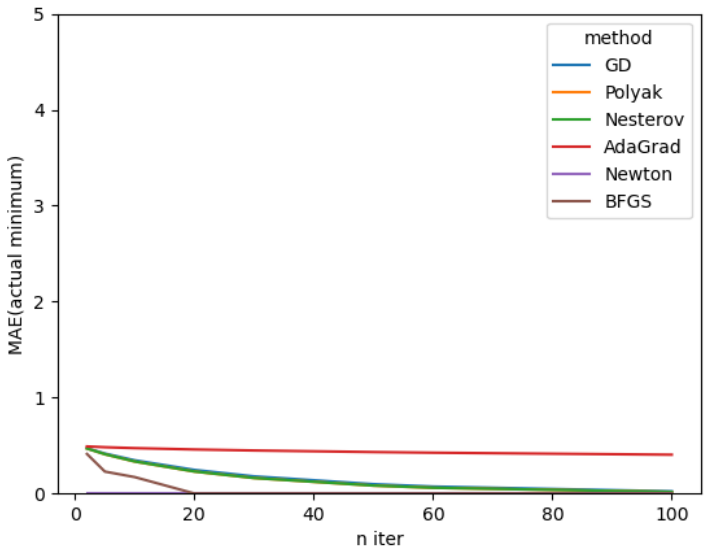

Figure 1: **Convergence on function (a).** All function converge almost imidiately. GD, Polyak and Nesterov converge relative to other function to the same point, roughly in the same time.

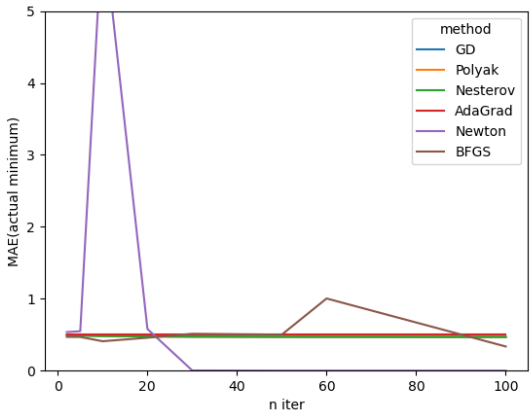

Figure 2: **Convergence on function (b).** Newton the best other functions. GD, Polyak and Nesterov converge relative to other function to the same point, roughly in the same time. It appear that given enough iterations BFGS is better than first order derivatives.

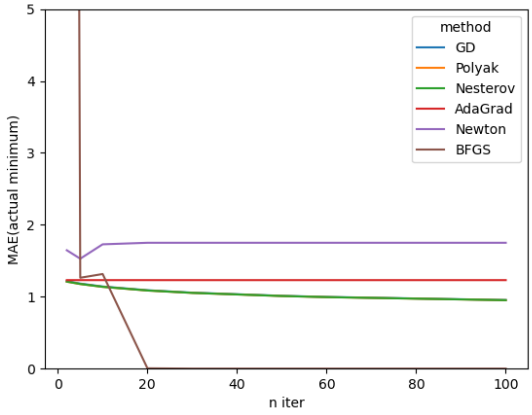

Figure 3: **Convergence on function (c).** First order derivative functions with the exception of AdaGrad are the most stable. Newton does not converge to minimum in either of the point and BFGS fails to converge at all in point [4.5, 4.5].

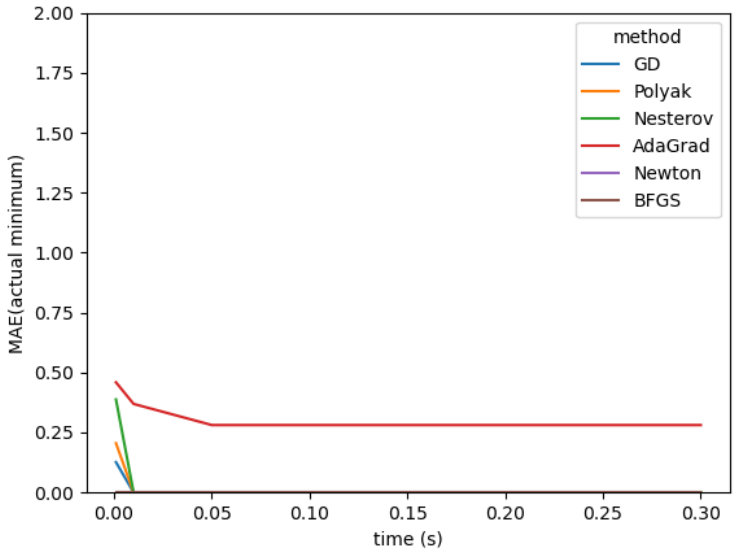

Figure 4: **Convergence on function (a).** All function converge almost imidiately to the minimum. At the very beginning we see different convergence rates for GD, Polyak and Nesterov.

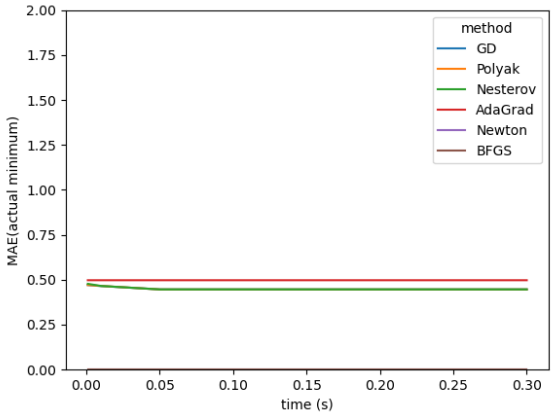

Figure 5: **Convergence on function (b).** Newton and BFGS second manage to find the minimum, while our outher methods fail. AdaGrad converges the most slowly.

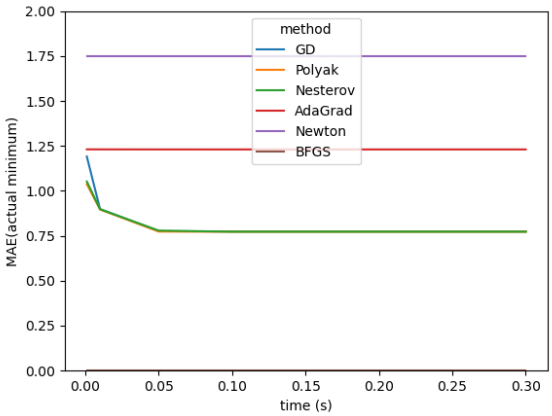

Figure 6: **Convergence on function (c).** First order derivative functions with the exception of AdaGrad are the most stable. Newton does not converge to minimum in either of the point and BFGS fails to converge in point [4.5, 4.5].

4 Linear Regression

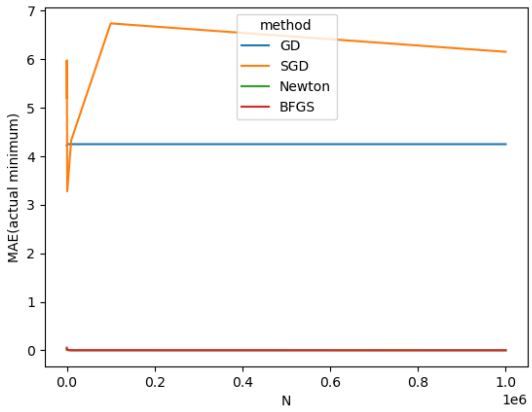

Figure 7: **MAE given 10 iteration steps.** Newton and BFGS converge in less than 10 iterations.

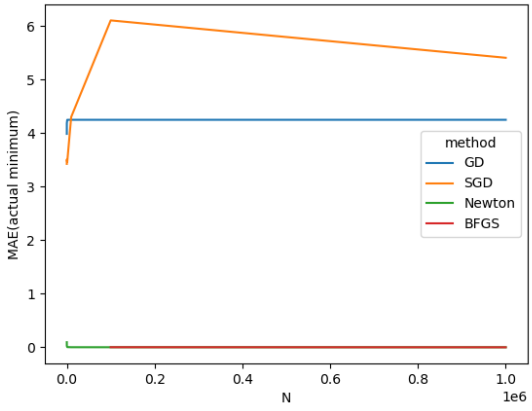

Figure 8: **MAE given 100 iteration steps.** Newton and BFGS converge quickly. while SGD appear to underperform GD.

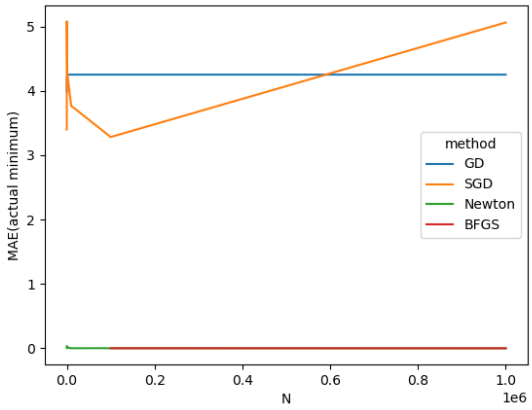

Figure 9: **MAE given 0.001 seconds.** SGD is much faster than GD, but slower than BFGS and Newton.

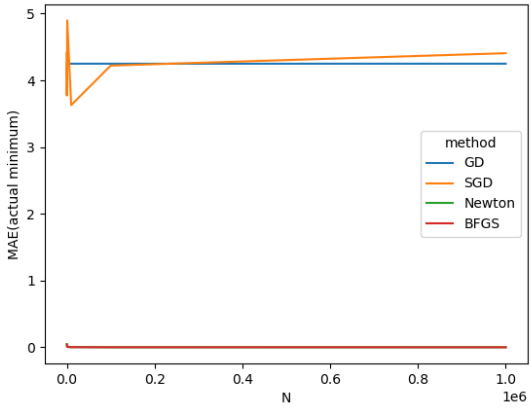

Figure 10: **MAE given 2 seconds.** While SGD performs better at first it later underperforms and appears to diverge from GD. BFGS and SGD converge quickly