Computing rho for three examples

Michael F. Zimmer

January 1, 2021

This document is meant to support the Neograd paper [Zimmer, 2020] by giving the step-by-step calculation of ρ for the three intro examples (quadratic, quartic, ellipse). In all the cases, gradient descent (GD) is used as the update scheme. Also, in the quadratic and quartic cases, the parameter $\boldsymbol{\theta}$ is allowed to be n-dimensional; for the ellipse, it is 2-dimensional.

1 Quadratic cost function

In this case the cost function is $f = c\boldsymbol{\theta}^2$. For basic GD, with $\alpha, c > 0$,

$$\begin{aligned} d\boldsymbol{\theta} &= -\alpha \boldsymbol{\nabla} f = -\alpha(2c\boldsymbol{\theta}) \\ f_{old} &= c\boldsymbol{\theta}^2 \\ f_{new} &= c(\boldsymbol{\theta} + d\boldsymbol{\theta})^2 \\ f_{est} &= f_{old} + \boldsymbol{\nabla} f \cdot d\boldsymbol{\theta} = c\boldsymbol{\theta}^2 + 2c\boldsymbol{\theta} \cdot d\boldsymbol{\theta} \end{aligned}$$

1.1 Computing ρ

Here we have

$$\begin{aligned} \rho &= \left| \frac{f_{new} - f_{est}}{f_{old} - f_{est}} \right| \\ &= \left| \frac{c(\boldsymbol{\theta} + d\boldsymbol{\theta})^2 - (c\boldsymbol{\theta}^2 + 2c\boldsymbol{\theta} \cdot d\boldsymbol{\theta})}{c\boldsymbol{\theta}^2 - (c\boldsymbol{\theta}^2 + 2c\boldsymbol{\theta} \cdot d\boldsymbol{\theta})} \right| \\ &= \left| \frac{c(d\boldsymbol{\theta})^2}{-2c\boldsymbol{\theta} \cdot d\boldsymbol{\theta}} \right| \\ &= \left| \frac{c(-2\alpha c\boldsymbol{\theta})^2}{-2c\boldsymbol{\theta} \cdot (-2\alpha c\boldsymbol{\theta})} \right| \\ &= \left| \frac{4\alpha^2 c^3 \boldsymbol{\theta}^2}{4\alpha c^2 \boldsymbol{\theta}^2} \right| \\ &= \alpha c \end{aligned}$$

where in the final line use was made of the requirement that both α, c are greater than 0. Also, this is the first example of a formula for determining α as a function of ρ:

$$\alpha = \rho/c$$

2 Quartic cost function

In this case the cost function is $f = c\boldsymbol{\theta}^4 \equiv c(\boldsymbol{\theta}\cdot\boldsymbol{\theta})^2$. For basic GD, with $\alpha, c > 0$,

$$\begin{aligned}
d\boldsymbol{\theta} &= -\alpha\boldsymbol{\nabla} f = -\alpha(4c(\boldsymbol{\theta}\cdot\boldsymbol{\theta})\boldsymbol{\theta}) = -4\alpha c\boldsymbol{\theta}^2\boldsymbol{\theta} \\
f_{old} &= c\boldsymbol{\theta}^4 \\
f_{new} &= c(\boldsymbol{\theta} + d\boldsymbol{\theta})^4 \\
f_{est} &= f_{old} + \boldsymbol{\nabla} f\cdot d\boldsymbol{\theta} = c\boldsymbol{\theta}^4 + (4c\boldsymbol{\theta}^2\boldsymbol{\theta})\cdot d\boldsymbol{\theta}
\end{aligned}$$

2.1 Computing ρ

Substituting in for these quantities, ρ can be computed as

$$\begin{aligned}
\rho &= \left|\frac{f_{new} - f_{est}}{f_{old} - f_{est}}\right| \\
&= \left|\frac{c(\boldsymbol{\theta} + d\boldsymbol{\theta})^4 - [c\boldsymbol{\theta}^4 + 4c\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot d\boldsymbol{\theta}]}{c\boldsymbol{\theta}^4 - [c\boldsymbol{\theta}^4 + 4c\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot d\boldsymbol{\theta}]}\right| \\
&= \left|\frac{(\boldsymbol{\theta} + d\boldsymbol{\theta})^4 - [\boldsymbol{\theta}^4 + 4\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot d\boldsymbol{\theta}]}{-4\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot d\boldsymbol{\theta}}\right|
\end{aligned}$$

where

$$\begin{aligned}
(\boldsymbol{\theta} + d\boldsymbol{\theta})^4 &= \{\boldsymbol{\theta}^2 + 2\boldsymbol{\theta}\cdot d\boldsymbol{\theta} + (d\boldsymbol{\theta})^2\}^2 \\
&= \{\boldsymbol{\theta}^2 + 2\boldsymbol{\theta}\cdot(-4\alpha c\boldsymbol{\theta}^2\boldsymbol{\theta}) + (-4\alpha c\boldsymbol{\theta}^2\boldsymbol{\theta})^2\}^2 \\
&= \{\boldsymbol{\theta}^2 - 8\alpha c(\boldsymbol{\theta}^2)^2 + (4\alpha c)^2(\boldsymbol{\theta}^2)^3\}^2 \\
&= \boldsymbol{\theta}^4\{1 - 8\alpha c\boldsymbol{\theta}^2 + (4\alpha c)^2(\boldsymbol{\theta}^2)^2\}^2 \\
&= \boldsymbol{\theta}^4[1 - 2x + x^2]^2 \\
&= \boldsymbol{\theta}^4[1(1 - 2x + x^2) - 2x(1 - 2x + x^2) + x^2(1 - 2x + x^2)] \\
&= \boldsymbol{\theta}^4[1 - 2x + x^2 - 2x + 4x^2 - 2x^3 + x^2 - 2x^3 + x^4] \\
&= \boldsymbol{\theta}^4[1 - 4x + 6x^2 - 4x^3 + x^4]
\end{aligned}$$

where $x = 4\alpha c\boldsymbol{\theta}^2$. Likewise,

$$\begin{aligned}
\boldsymbol{\theta}^4 + 4\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot d\boldsymbol{\theta} &= \boldsymbol{\theta}^4 + 4\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot(-4\alpha c\boldsymbol{\theta}^2\boldsymbol{\theta}) \\
&= \boldsymbol{\theta}^4 - 16\alpha c(\boldsymbol{\theta}^2)^3 \\
&= \boldsymbol{\theta}^4[1 - 16\alpha c\boldsymbol{\theta}^2] \\
&= \boldsymbol{\theta}^4(1 - 4x)
\end{aligned}$$

Also,

$$\begin{aligned}
-4\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot d\boldsymbol{\theta} &= -4\boldsymbol{\theta}^2\boldsymbol{\theta}\cdot(-4\alpha c\boldsymbol{\theta}^2\boldsymbol{\theta}) \\
&= 16\alpha c(\boldsymbol{\theta}^2)^3 \\
&= 4(4\alpha c\boldsymbol{\theta}^2)\boldsymbol{\theta}^4 \\
&= 4x\boldsymbol{\theta}^4
\end{aligned}$$

It follows that

$$\begin{aligned}
\text{numerator} &= \boldsymbol{\theta}^4[1-4x+6x^2-4x^3+x^4] - \boldsymbol{\theta}^4(1-4x) \\
&= \boldsymbol{\theta}^4[6x^2-4x^3+x^4] \\
\text{denominator} &= \boldsymbol{\theta}^4[4x]
\end{aligned}$$

Reassembling...

$$\begin{aligned}
\rho &= \left|\frac{\boldsymbol{\theta}^4[6x^2-4x^3+x^4]}{\boldsymbol{\theta}^4[4x]}\right| \\
&= \left|\frac{6x-4x^2+x^3}{4}\right|
\end{aligned}$$

Define $x = 4q$, so that $q = \alpha c\boldsymbol{\theta}^2$.

$$\begin{aligned}
6x-4x^2+x^3 &= 24q - 4^3q^2 + 4^3q^3 \\
&= 4(6q-16q^2+16q^3)
\end{aligned}$$

and finally

$$\rho = \left|6q-16q^2+16q^3\right|$$

Note that for small q, ρ

$$\begin{aligned}
\rho &\approx 6q = 6\alpha c\boldsymbol{\theta}^2 \\
\alpha &\approx \rho/(6c\boldsymbol{\theta}^2)
\end{aligned}$$

3 Ellipse cost function

In this case there is a 2-dimensional parameter: $\boldsymbol{\theta} = (\theta_1, \theta_2)$

$$\begin{aligned}
f &= \frac{\theta_1^2}{a^2} + \frac{\theta_2^2}{b^2} \\
d\boldsymbol{\theta} &= -\alpha\boldsymbol{\nabla} f = -\alpha\left(\frac{2\theta_1}{a^2}, \frac{2\theta_2}{b^2}\right)
\end{aligned}$$

$$
\begin{aligned}
f_{old} &= f(\boldsymbol{\theta}) \\
&= \frac{\theta_1^2}{a^2} + \frac{\theta_2^2}{b^2} \\
f_{new} &= f(\boldsymbol{\theta} + d\boldsymbol{\theta}) \\
&= \frac{1}{a^2}\left(\theta_1 - \frac{2\alpha\theta_1}{a^2}\right)^2 + \frac{1}{b^2}\left(\theta_2 - \frac{2\alpha\theta_2}{b^2}\right)^2 \\
&= \frac{1}{a^2}\left[\theta_1^2 - 2\theta_1\frac{2\alpha\theta_1}{a^2} + (\frac{2\alpha\theta_1}{a^2})^2\right] + \frac{1}{b^2}\left[\theta_2^2 - 2\theta_2\frac{2\alpha\theta_2}{b^2} + (\frac{2\alpha\theta_2}{b^2})^2\right] \\
&= f_{old} - \frac{4\alpha\theta_1^2}{a^4} + \frac{4\alpha^2\theta_1^2}{a^6} - \frac{4\alpha\theta_2^2}{b^4} + \frac{4\alpha^2\theta_2^2}{b^6} \\
f_{est} &= f_{old} + \boldsymbol{\nabla} f \cdot d\boldsymbol{\theta} \\
&= f_{old} - \alpha\left(\frac{2\theta_1}{a^2}, \frac{2\theta_2}{b^2}\right) \cdot \left(\frac{2\theta_1}{a^2}, \frac{2\theta_2}{b^2}\right) \\
&= f_{old} - \alpha\frac{4\theta_1^2}{a^4} - \alpha\frac{4\theta_2^2}{b^4}
\end{aligned}
$$

Also, define this

$$Q_m = \frac{\theta_1^2}{a^m} + \frac{\theta_2^2}{b^m}$$

Now can start putting the pieces together...

$$
\begin{aligned}
f_{new} - f_{est} &= \frac{4\alpha^2\theta_1^2}{a^6} + \frac{4\alpha^2\theta_2^2}{b^6} \\
f_{old} - f_{est} &= \alpha\frac{4\theta_1^2}{a^4} + \alpha\frac{4\theta_2^2}{b^4}
\end{aligned}
$$

ρ may be expressed as

$$
\begin{aligned}
\rho &= \left|\frac{f_{new} - f_{est}}{f_{old} - f_{est}}\right| \\
&= \left|\frac{4\alpha^2 Q_6}{4\alpha Q_4}\right| \\
&= \alpha\frac{Q_6}{Q_4}
\end{aligned}
$$

The absolute values can be removed since all the quantities in this expression are positive. This expression easily allows for the inversion:

$$\alpha = \rho\frac{Q_4}{Q_6}$$

This means that once a decision is made for the target value of ρ, α can be determined.

References

M.F. Zimmer. Neograd: Gradient descent with a near-ideal learning rate, 2020. URL http://www.arxiv.org/pdf/2010.07873.pdf.