Ridge Problem in Stochastic Gradient Descent: An Analytical Solution

Tran Phong Binh*

Department of Computer Science, National Tsing Hua University

July 15, 2021

Abstract

1 Introduction

2 Proof

$$f(x, y) = \begin{cases} \phi(x, y) & \text{if } x = 2y, \\ \psi(x, y) & \text{otherwise,} \end{cases} \tag{1}$$

where

$$\phi(x, y) = x^2 + y^2 - 8x - 4y + 10, \tag{2}$$

$$\psi(x, y) = x^2 + y^2 - 4x - 2y. \tag{3}$$

*Email: phongbinh2511@gmail.com

2.1   First Partial Derivatives

We examine the first derivatives of our function f at point $(2, 1)$.

$$\frac{\partial f}{\partial x}(2, 1) = \lim_{h \to 0} \frac{f(2 + h, 1) - f(2, 1)}{h} \tag{4}$$

Because the numerator

$$f(2 + h, 1) = \psi(2 + h, 1) \tag{5}$$

$$\begin{aligned} f(2, 1) &= \phi(2, 1) \\ &= -5 \\ &= \psi(2, 1) \end{aligned} \tag{6}$$

it holds that

$$\begin{aligned} \frac{\partial f}{\partial x}(2, 1) &= \frac{\partial \psi}{\partial x}(2, 1) \\ &= 2x - 4 \\ &= 0 \end{aligned} \tag{7}$$

Similarly,

$$\begin{aligned} \frac{\partial f}{\partial y}(2, 1) &= \frac{\partial \psi}{\partial y}(2, 1) \\ &= 2y - 2 \\ &= 0 \end{aligned} \tag{8}$$

Hence

$$\nabla_f(2, 1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{9}$$

2.2   Second Partial Derivatives

We then consider the second partial derivatives of our function at the same point.

2.2.1   Second Partial Derivative With Respect To x

We begin with $f_{xx}(2, 1)$:

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(2, 1) = \lim_{h \to 0} \frac{\frac{\partial f}{\partial x}(2 + h, 1) - \frac{\partial f}{\partial x}(2, 1)}{h} \tag{10}$$

Assessing the first component of the numerator

$$\frac{\partial f}{\partial x}(2+h,1) = \lim_{k \to 0} \frac{f(2+h+k,1) - f(2+h,1)}{k} \tag{11}$$

We observe that

$$f(2+h+k,1) = \begin{cases} \psi(2+h+k,1) & \text{if } h \to 0^-, k \to 0^- \\ \phi(2,1) = -5 = \psi(2,1) & \text{if } h \to 0^-, k \to 0^+ \\ \phi(2,1) = -5 = \psi(2,1) & \text{if } h \to 0^+, k \to 0^- \\ \psi(2+h+k,1) & \text{if } h \to 0^+, k \to 0^+ \end{cases} \tag{12}$$

$$= \psi(2+h+k,1) \tag{13}$$

and

$$f(2+h,1) = \psi(2+h,1) \tag{14}$$

Hence

$$\frac{\partial f}{\partial x}(2+h,1) = \frac{\partial \psi}{\partial x}(2+h,1) \tag{15}$$

By equations 7 and 15,

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(2,1) = \frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial x}\right)(2,1) \tag{16}$$

$$= 2 \tag{17}$$

2.2.2 Second Partial Derivative With Respect To x Then y

We continue by examining $f_{yx}(2,1)$:

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(2,1) = \lim_{h \to 0} \frac{\frac{\partial f}{\partial x}(2,1+h) - \frac{\partial f}{\partial x}(2,1)}{h} \tag{18}$$

Again, we assess the first component of the numerator

$$\frac{\partial f}{\partial x}(2,1+h) = \lim_{k \to 0} \frac{f(2+k,1+h) - f(2,1+h)}{k} \tag{19}$$

Similar to $f_{xx}(2,1)$, we observe that

$$f(2+k,1+h) = \begin{cases} \psi(2+k,1+h) & \text{if } h \to 0^-, k \to 0^- \\ \psi(2+k,1+h) & \text{if } h \to 0^-, k \to 0^+ \\ \psi(2+k,1+h) & \text{if } h \to 0^+, k \to 0^- \\ \psi(2+k,1+h) & \text{if } h \to 0^+, k \to 0^+ \end{cases} \tag{20}$$

$$= \psi(2+k,1+h) \tag{21}$$

3

and
$$f(2, 1+h) = \psi(2, 1+h) \tag{22}$$

Hence
$$\frac{\partial f}{\partial x}(2, 1+h) = \frac{\partial \psi}{\partial x}(2, 1+h) \tag{23}$$

By equations 7 and 23,

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(2, 1) = \frac{\partial}{\partial y}\left(\frac{\partial \psi}{\partial x}\right)(2, 1) \tag{24}$$
$$= 0 \tag{25}$$

2.2.3   Hessian Determinant

Performing similar deductions for $f_{xy}(2, 1)$ and $f_{yy}(2, 1)$, we have

$$f_{xx}(2, 1) = 2 \tag{26}$$
$$f_{yx}(2, 1) = 0 \tag{27}$$
$$f_{xy}(2, 1) = 0 \tag{28}$$
$$f_{yy}(2, 1) = 2 \tag{29}$$

Hence the Hessian determinant

$$H_f = f_{xx}(2, 1)\, f_{yy}(2, 1) - f_{yx}(2, 1)\, f_{xy}(2, 1) \tag{30}$$
$$= 4 > 0 \tag{31}$$

2.3   Second Partial Derivative Test Contradiction

According to [?], by equations 9, 31, and the fact that $f_{xx}(2, 1) = 2 > 0$, we shall conclude $(2, 1)$ is a local minimum point. However, this is incorrect, as for $h \in \mathbb{R}, h \to 0^+$:

$$f(2+2h, 1+h) = \phi(2+2h, 1+h) \tag{32}$$
$$< \phi(2, 1) = f(2, 1) \tag{33}$$

That is, $(2, 1)$ is not a local minimum point (proof in appendix).

3 Intuition

4 Solution

For the tangent hyperplane of a multivariable function to be 'flat', it is not enough for the gradient of the function at that point to be the zero vector, but the first derivatives of the function in all directions at the point of interest must equal to 0.

In this work, we provide an analytical solution to this issue by redefining what it takes for the tangent hyperplane of a function at a point to be 'flat':

Definition 4.1. *Given a multivariable function*

$$f \colon \mathbb{R}^{\mathbb{N}} \longrightarrow \mathbb{R}$$
$$\mathbf{x} \longmapsto f(\mathbf{x})$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. *The tangent hyperplane of the function is flat at point* $\mathbf{x_0}$ *iff for all directions* $\mathbf{v} \in \mathbb{R}^{\mathbb{N}}$, *the single variable function*

$$g \colon \mathbb{R} \longrightarrow \mathbb{R}$$
$$k \longmapsto f(\mathbf{x_0} + k\mathbf{v})$$

has zero derivative at $k = 0$ *i.e. iff* $\forall \mathbf{v} \in \mathbb{R}^{\mathbb{N}} \colon \frac{\mathrm{d}g}{\mathrm{d}k}(0) = 0$.

5 Conclusion

6 Future Work