Ridge Problem in Stochastic Gradient Descent: An Analytical Solution

Tran Phong Binh*

Department of Computer Science, National Tsing Hua University

July 19, 2021

Abstract

1 Introduction

2 Proof

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

where

$$\phi(x,y) = x^2 + y^2 - 8x - 4y + 10,$$
$$\psi(x,y) = x^2 + y^2 - 4x - 2y.$$

*Email: phongbinh2511@gmail.com

2.1 First Partial Derivatives

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

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

Because the numerator

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

$$\begin{aligned} f(2,1) &= \phi(2,1) \\ &= -5 \\ &= \psi(2,1) \end{aligned} \tag{3}$$

it holds that

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

Similarly,

$$\begin{aligned} \frac{\partial f}{\partial y}(2,1) &= \frac{\partial \psi}{\partial y}(2,1) \\ &= 2y - 2 \\ &= 0 \end{aligned} \tag{5}$$

Hence

$$\nabla_f(2,1) = \begin{bmatrix} 0 \\ 0 \end{bmatrix} \tag{6}$$

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 $f_{xx}(2,1)$:

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(2,1) = \lim_{h\to 0}\frac{\frac{\partial f}{\partial x}(2+h,1) - \frac{\partial f}{\partial x}(2,1)}{h} \tag{7}$$

Assessing the first component of the numerator

$$\frac{\partial f}{\partial x}(2+h,1) = \lim_{k\to 0}\frac{f(2+h+k,1) - f(2+h,1)}{k} \tag{8}$$

We observe that

$$f(2+h+k,1) = \begin{cases} \psi(2+h+k,1) & \text{if } h\to 0^-,\ k\to 0^- \\ \phi(2,1) = -5 = \psi(2,1) & \text{if } h\to 0^-,\ k\to 0^+ \\ \phi(2,1) = -5 = \psi(2,1) & \text{if } h\to 0^+,\ k\to 0^- \\ \psi(2+h+k,1) & \text{if } h\to 0^+,\ k\to 0^+ \end{cases} \tag{9}$$

$$= \psi(2+h+k,1) \tag{10}$$

and

$$f(2+h,1) = \psi(2+h,1) \tag{11}$$

Hence

$$\frac{\partial f}{\partial x}(2+h,1) = \frac{\partial \psi}{\partial x}(2+h,1) \tag{12}$$

By equations 4 and 12,

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(2,1) = \frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial x}\right)(2,1) \tag{13}$$

$$= 2 \tag{14}$$

2.2.2 Second Partial Derivative With Respect To x Then y

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

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(2,1) = \lim_{h\to 0}\frac{\frac{\partial f}{\partial x}(2,1+h) - \frac{\partial f}{\partial x}(2,1)}{h} \tag{15}$$

Again, we assess the first component of the numerator

$$\frac{\partial f}{\partial x}(2,1+h) = \lim_{k\to 0}\frac{f(2+k,1+h) - f(2,1+h)}{k} \tag{16}$$

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

$$f(2+k,1+h) = \begin{cases} \psi(2+k,1+h) & \text{if } h \to 0^-,\ k \to 0^- \\ \psi(2+k,1+h) & \text{if } h \to 0^-,\ k \to 0^+ \\ \psi(2+k,1+h) & \text{if } h \to 0^+,\ k \to 0^- \\ \psi(2+k,1+h) & \text{if } h \to 0^+,\ k \to 0^+ \end{cases} \tag{17}$$

$$= \psi(2+k,1+h) \tag{18}$$

and

$$f(2,1+h) = \psi(2,1+h) \tag{19}$$

Hence

$$\frac{\partial f}{\partial x}(2,1+h) = \frac{\partial \psi}{\partial x}(2,1+h) \tag{20}$$

By equations 4 and 20,

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(2,1) = \frac{\partial}{\partial y}\left(\frac{\partial \psi}{\partial x}\right)(2,1) \tag{21}$$

$$= 0 \tag{22}$$

2.2.3 Hessian Determinant

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

$$f_{xx}(2,1) = 2 \tag{23}$$

$$f_{yx}(2,1) = 0 \tag{24}$$

$$f_{xy}(2,1) = 0 \tag{25}$$

$$f_{yy}(2,1) = 2 \tag{26}$$

Hence the Hessian determinant

$$H_f = f_{xx}(2,1) f_{yy}(2,1) - f_{yx}(2,1) f_{xy}(2,1) \tag{27}$$

$$= 4 > 0 \tag{28}$$

2.3 Second Partial Derivative Test Contradiction

According to [1], by equations 6, 28, 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 $h \in \mathbb{R}$, $h \to 0^+$:

$$f(2+2h, 1+h) = \phi(2+2h, 1+h) \tag{29}$$
$$< \phi(2,1) = f(2,1) \tag{30}$$

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. Firstly, we define the concept of directional function[1]:

Definition 4.1 *Given a multivariable function*

$$\begin{aligned} f\colon \mathbb{R}^n &\longrightarrow \mathbb{R} \\ \mathbf{x} &\longmapsto f(\mathbf{x}) \end{aligned}$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$. *The directional function of* f *at point* $\mathbf{x_0}$ *in direction* $\mathbf{v} \in \mathbb{R}^n$ *is the single variable function*

$$\begin{aligned} \xi_{\mathbf{v}}\colon \mathbb{R} &\longrightarrow \mathbb{R} \\ k &\longmapsto f(\mathbf{x_0} + k\mathbf{v}) \end{aligned}$$

With this, we are ready to redefine flat tangent hyperplane:

Definition 4.2 *Given a multivariable function* f*, its tangent hyperplane is flat at point* $\mathbf{x_0}$ *iff for all directions* $\mathbf{v} \in \mathbb{R}^n$*, the directional function* $\xi_{\mathbf{v}}$ *at that point has zero first derivative at* $k = 0$ *i.e. iff* $\forall \mathbf{v} \in \mathbb{R}^n\colon \frac{\mathrm{d}\xi_{\mathbf{v}}}{\mathrm{d}k}(0) = 0$.

From this definition, we can derive our new local mininimum test:

[1] We exclude the vector $\mathbf{0}$ in all of our discussions on directional function.

Theorem 4.1 *Given a multivariable function f, if its tangent hyperplane is flat at point $\mathbf{x_0}$, and for all directions $\mathbf{v} \in \mathbb{R}^n$, the directional function $\xi_\mathbf{v}$ at that point has positive second derivative at $k = 0$ i.e. if $\forall \mathbf{v} \in \mathbb{R}^n \colon \frac{\mathrm{d}\xi_\mathbf{v}}{\mathrm{d}k}(0) = 0 \wedge \frac{\mathrm{d}^2\xi_\mathbf{v}}{\mathrm{d}k^2}(0) > 0$, then $\mathbf{x_0}$ is a local minimum point.*

4.1 False Local Minimum

We revisit our first example to see if the test works:

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

where

$$\begin{aligned} \phi(x,y) &= x^2 + y^2 - 8x - 4y + 10, \\ \psi(x,y) &= x^2 + y^2 - 4x - 2y. \end{aligned}$$

For $\mathbf{v} = (4,2)$, our directional function at $\mathbf{x_0} = (2,1)$ is

$$\begin{aligned} \xi_\mathbf{v}(k) &= f(\mathbf{x_0} + k\mathbf{v}) && (31) \\ &= f\left(\begin{bmatrix}2\\1\end{bmatrix} + k\begin{bmatrix}4\\2\end{bmatrix}\right) && (32) \\ &= f\left(\begin{bmatrix}2+4k\\1+2k\end{bmatrix}\right) \text{ or } f(2+4k, 1+2k) && (33) \end{aligned}$$

Since $x = 2y$,

$$\begin{aligned} \xi_\mathbf{v}(k) &= \phi(2+4k, 1+2k) && (34) \\ &= (2+4k)^2 + (1+2k)^2 - 8(2+4k) - 4(1+2k) + 10 && (35) \\ &= 4 + 16k + 16k^2 + 1 + 4k + 4k^2 - 16 - 32k - 4 - 8k + 10 && (36) \\ &= 20k^2 - 20k - 5 && (37) \end{aligned}$$

We examine the directional function's first derivative at $k = 0$:

$$\begin{aligned} \frac{\mathrm{d}\xi_\mathbf{v}}{\mathrm{d}k}(0) &= 40k - 20 && (38) \\ &= -20 \neq 0 && (39) \end{aligned}$$

Hence $\mathbf{x_0} = (2,1)$ is not a local minimum point!

4.2 True Local Minimum

We consolidate our test with the well-known function

$$f(\mathbf{x}) = \mathbf{x}^\intercal \mathrm{A} \mathbf{x} \tag{40}$$

where $\underset{n \times n}{\mathrm{A}} \succ 0$. For any $\mathbf{v} \in \mathbb{R}^n$, our directional function at $\mathbf{x_0} = \mathbf{0}$ is

$$\begin{aligned} \xi_{\mathbf{v}}(k) &= f(\mathbf{x_0} + k\mathbf{v}) && (41) \\ &= (k\mathbf{v})^\intercal \mathrm{A} (k\mathbf{v}) && (42) \\ &= k^2 \mathbf{v}^\intercal \mathrm{A} \mathbf{v} && (43) \end{aligned}$$

We examine the directional function's first derivative at $k = 0$:

$$\begin{aligned} \frac{\mathrm{d}\xi_{\mathbf{v}}}{\mathrm{d}k}(0) &= 2k\mathbf{v}^\intercal \mathrm{A} \mathbf{v} && (44) \\ &= 0 && (45) \end{aligned}$$

Furthermore,

$$\frac{\mathrm{d}^2 \xi_{\mathbf{v}}}{\mathrm{d}k^2}(0) = 2\mathbf{v}^\intercal \mathrm{A} \mathbf{v} \tag{46}$$

Since $A \succ 0$, $\forall \mathbf{v} \neq \mathbf{0}$: $\mathbf{v}^\intercal \mathrm{A} \mathbf{v} > 0$. Thus, our directional function's second derivative is positive. According to Theorem 4.1, $\mathbf{x_0} = \mathbf{0}$ is a local minimum point!

5 Conclusion

6 Future Work

References

[1] Eric W. Weisstein. *Second Derivative Test*. URL: `https://mathworld.wolfram.com/SecondDerivativeTest.html` (visited on 07/15/2021).