Ridge Problem in Stochastic Gradient Descent: An Analytical Solution

Phong-Binh Tran*

Department of Computer Science, National Tsing Hua University

September 2, 2022

Abstract

1 Introduction

The second partial derivative test is a renowned technique for determining local extremum points. This mathematical method had been well studied since the beginning era of calculus, and has been used extensively both in the academia and in real-world applications. Yet, upon studying the ridge problem of hill climbing[2], we stumbled on a thought that the same issue could present in the forementioned test under a stochastic space setup.

In this work, we prove that such problem exists, proposing an analytical solution, and check our approach on two multivariable functions. We hope our finding further pushes knowledge boundaries of stochastic gradient descent in specific, and optimization in general. With this, researchers and practitioners can now surpass false local extremum points to navigate down true local optimization objectives, consolidating the robustness of their theories and systems.

*Email: phongbinh2511@gmail.com

2 Proof

We prove that ridge problem presents in the second partial derivative test by analyzing the multivariable function

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

where

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

at point $(2,1)$.

2.1 First Partial Derivatives

We begin by examining the first derivatives of f at point $(2,1)$. Firstly, it is given that

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

Because the components on the numerator are

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

it holds that

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

Similarly,

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

Hence

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

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

$$\frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(2,1) = \lim_{h\to 0} \frac{\frac{\partial f}{\partial x}(2+h,1) - \frac{\partial f}{\partial x}(2,1)}{h}. \tag{3}$$

Assessing the first component of the numerator

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

we observe that

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

and

$$f(2+h,1) = \psi(2+h,1).$$

Therefore, our first entry of the forementioned numerator

$$\frac{\partial f}{\partial x}(2+h,1) = \frac{\partial \psi}{\partial x}(2+h,1). \tag{4}$$

By equations 1 and 4, our equation 3 becomes

$$\begin{aligned} \frac{\partial}{\partial x}\left(\frac{\partial f}{\partial x}\right)(2,1) &= \frac{\partial}{\partial x}\left(\frac{\partial \psi}{\partial x}\right)(2,1) \\ &= 2. \end{aligned}$$

2.2.2 Second Partial Derivative With Respect To x Then y

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

$$\frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(2,1) = \lim_{h\to 0}\frac{\frac{\partial f}{\partial x}(2,1+h) - \frac{\partial f}{\partial x}(2,1)}{h}. \tag{5}$$

Again, we assess the first component of the numerator

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

Analogous to $f_{xx}(2,1)$, we learn that

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

and

$$f(2,1+h) = \psi(2,1+h).$$

Thus

$$\frac{\partial f}{\partial x}(2,1+h) = \frac{\partial \psi}{\partial x}(2,1+h). \tag{6}$$

By equations 1 and 6, our equation 5 becomes

$$\begin{aligned} \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)(2,1) &= \frac{\partial}{\partial y}\left(\frac{\partial \psi}{\partial x}\right)(2,1) \\ &= 0. \end{aligned}$$

2.2.3 Hessian Matrix

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

$$\begin{aligned} f_{xx}(2,1) &= 2, \\ f_{yx}(2,1) &= 0, \\ f_{xy}(2,1) &= 0, \\ f_{yy}(2,1) &= 2. \end{aligned}$$

As a result, the Hessian determinant at point $(2, 1)$

$$\begin{aligned} |\text{Hess } f_{(2,1)}| &= f_{xx}(2,1) f_{yy}(2,1) - f_{yx}(2,1) f_{xy}(2,1) \\ &= 4. \end{aligned}$$

Since $f_{xx}(2,1) > 0$ and $|\text{Hess } f_{(2,1)}| > 0$,

$$\text{Hess } f_{(2,1)} \succ 0. \tag{7}$$

2.3 Second Partial Derivative Test Contradiction

According to [1], by equations 2 and 7, we shall conclude $(2, 1)$ is a local minimum point. However, this is incorrect, as for $h \in \mathbb{R}$, $h \to 0^+$:

$$f(2 + 2h, 1 + h) = \phi(2 + 2h, 1 + h) \tag{8}$$

$$< \phi(2, 1) = f(2, 1) \tag{9}$$

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

3 Intuition

Considering our function's graph depicted in Figure 1 at point $(2, 1)$, we observe that the first derivatives in the x and y directions are zeroes, making the zero gradient. Furthermore, the curvature of the function is up in all directions at the point of interest, indicating the positive definite Hessian. With this premise, the second partial derivative test concludes that $(2, 1)$ is a local minimum point.

However, this is analytically and visually false. Even though the curvature along the $\mathbf{v} = (4, 2)$ direction is up, the corresponding first derivative differs from zero, invalidating the second partial derivative test.

4 Solution

For the tangent hyperplane of a multivariable function at a point 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 at the point of interest must equal to zero in all directions.

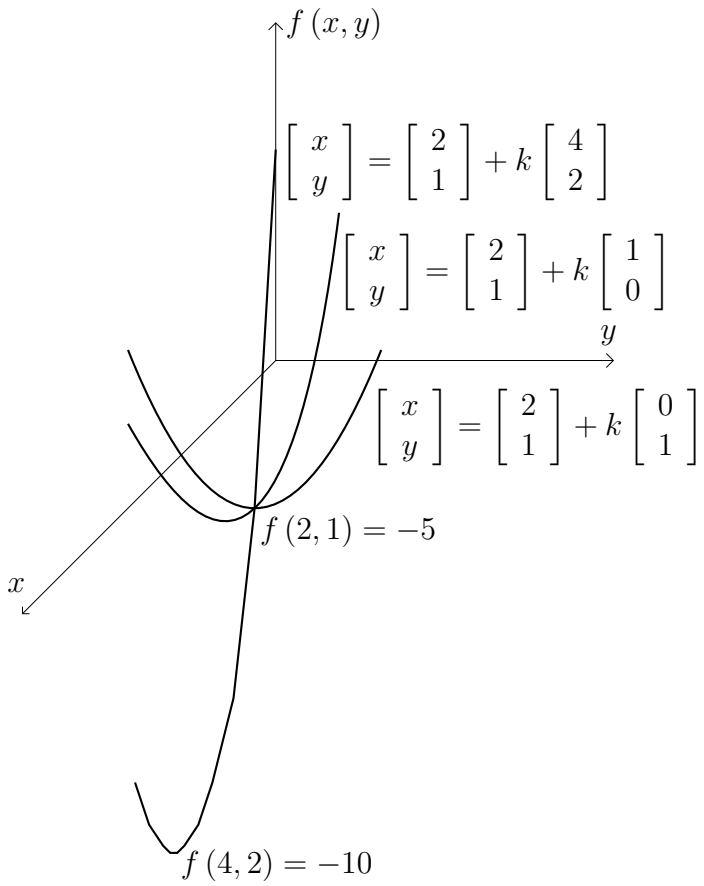

Figure 1: Partial graph of the function $f(x,y)$

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:

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 $\mathbf{x} = (x_1, x_2, \ldots, x_n)$*. For a direction*[1] $\mathbf{v} \in \mathbb{R}^n$*, the directional function*

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

of f at point $\mathbf{x_0}$ is the single variable function

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

We are now 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 $\forall \mathbf{v} \in \mathbb{R}^n : \frac{\mathrm{d}\xi_{\mathbf{v}}}{\mathrm{d}k}(0) = 0$.*

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

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 $\forall \mathbf{v} \in \mathbb{R}^n : \frac{\mathrm{d}\xi_{\mathbf{v}}}{\mathrm{d}k}(0) = 0 \wedge \frac{\mathrm{d}^2\xi_{\mathbf{v}}}{\mathrm{d}k^2}(0) > 0$, 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 the direction $\mathbf{v} = (4, 2)$, our directional function at $\mathbf{x_0} = (2, 1)$ is

$$\begin{aligned}\xi_{\mathbf{v}}(k) &= f(\mathbf{x_0} + k\mathbf{v}) && (10) \\ &= f\left(\begin{bmatrix}2\\1\end{bmatrix} + k\begin{bmatrix}4\\2\end{bmatrix}\right) && (11) \\ &= f\left(\begin{bmatrix}2+4k\\1+2k\end{bmatrix}\right) \text{ or } f(2+4k, 1+2k) && (12)\end{aligned}$$

Since $x = 2y$,

$$\xi_{\mathbf{v}}(k) = \phi(2 + 4k, 1 + 2k) \tag{13}$$
$$= (2 + 4k)^2 + (1 + 2k)^2 - 8(2 + 4k) - 4(1 + 2k) + 10 \tag{14}$$
$$= 4 + 16k + 16k^2 + 1 + 4k + 4k^2 - 16 - 32k - 4 - 8k + 10 \tag{15}$$
$$= 20k^2 - 20k - 5 \tag{16}$$

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

$$\frac{\mathrm{d}\xi_{\mathbf{v}}}{\mathrm{d}k}(0) = 40k - 20 \tag{17}$$
$$= -20 \neq 0 \tag{18}$$

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

$$f(\mathbf{x}) = \mathbf{x}^\intercal \mathrm{A} \mathbf{x} \tag{19}$$

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

$$\xi_{\mathbf{v}}(k) = f(\mathbf{x_0} + k\mathbf{v}) \tag{20}$$
$$= (k\mathbf{v})^\intercal \mathrm{A} (k\mathbf{v}) \tag{21}$$
$$= k^2 \mathbf{v}^\intercal \mathrm{A} \mathbf{v} \tag{22}$$

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

$$\frac{\mathrm{d}\xi_{\mathbf{v}}}{\mathrm{d}k}(0) = 2k\mathbf{v}^\intercal \mathrm{A} \mathbf{v} \tag{23}$$
$$= 0 \tag{24}$$

Furthermore,

$$\frac{\mathrm{d}^2\xi_{\mathbf{v}}}{\mathrm{d}k^2}(0) = 2\mathbf{v}^\intercal \mathrm{A} \mathbf{v} \tag{25}$$

Since $\mathrm{A} \succ 0$, $\forall \mathbf{v} \neq \mathbf{0} \colon \mathbf{v}^\intercal \mathrm{A} \mathbf{v} > 0$. 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).

[2] Patrick Henry Winston. *Artificial Intelligence (3rd Edition)*. Pearson, 1992, pp. 73–74. ISBN: 0201533774.