Deep learning by Ian goodfellow

- Chapter 4 - Numerical Computation
  - 4.1 - Overflow and Underflow
    - What is rounding error?

      Rounding error is problematic, especially when it compounds across many operations, and can cause algorithms that work in theory to fail in practice if they are not designed to minimize the accumulation of rounding error.

      One form of rounding error that is particularly devastating isunderflow.Underflow occurs when numbers near zero are rounded to zero. Many functionsbehave qualitatively differently when their argument is zero rather than a smallpositive number. For example, we usually want to avoid division by zero.

      Another highly damaging form of numerical error isoverflow. Overflow occurs when numbers with large magnitude are approximated as ∞ or −∞.

    - What f(x) helps to prevent rounding error?

      One example of a function that must be stabilized against underflow and overflow is the softmax function. The softmax function is often used to predict the probabilities associated with a multinoulli distribution. The softmax function is defined to be

$$\text{softmax}(\boldsymbol{x})_i = \frac{\exp(x_i)}{\sum_{j=1}^{n} \exp(x_j)}.$$

    - What is the downside of softmax function?

      Underflow in the numerator can still cause the expression as a whole to evaluate to zero. This means that if we implementlog softmax(x) by first running the softmax subroutine then passing the result to the log function, we could erroneously obtain**−∞.** Instead, we must implement a separate function that calculates log softmaxin a numerically stable way. The log

softmax function can be stabilized using the same trick as we used to stabilize the softmax function.

▼ 4.2 - Poor Conditioning

▼ What does conditioning refer to?

Conditioning refers to how rapidly a function changes with respect to small changes in its inputs. Functions that change rapidly when their inputs are perturbed slightly can be problematic for scientific computation because rounding errors in the input scan result in large changes in the output.

▼ 4.3 - Gradient-Based Optimization

▼ What is an objective function?

The function we want to minimize or maximize is called the objective function, or criterion. When we are minimizing it, we may also call it the cost function, loss function, or error function.

▼ What is gradient descent?

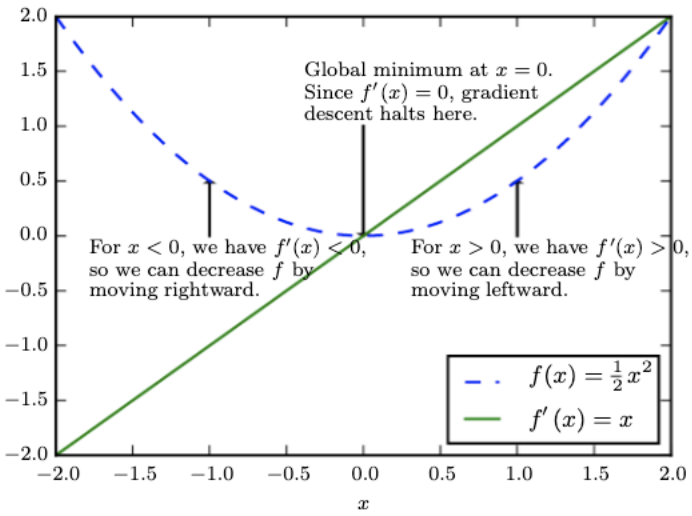

Figure 4.1: Gradient descent. An illustration of how the gradient descent algorithm uses the derivatives of a function to follow the function downhill to a minimum.

Suppose we have a functiony=f(x), where both x and y are real numbers. The derivative of this function is denoted as f□(x) or as dydx. The derivative f□(x)gives the slope off(x) at the point x. In other words, it specifies how to

scale a small change in the input to obtain the corresponding change in the output:f(x + □) ≈ f(x) + □f□(x).

The derivative is therefore useful for minimizing a function because it tells us how to change x in order to make a small improvement in y. For example, we know that f(x − □ sign(f□(x))) is less than f(x) for small enough□. We can thus reduce f(x) by moving x in small steps with the opposite sign of the derivative.This technique is called gradient descent(Cauchy, 1847). See figure 4.1 for an example of this technique

Points wheref□(x) = 0 are known as critical points, or stationary points.

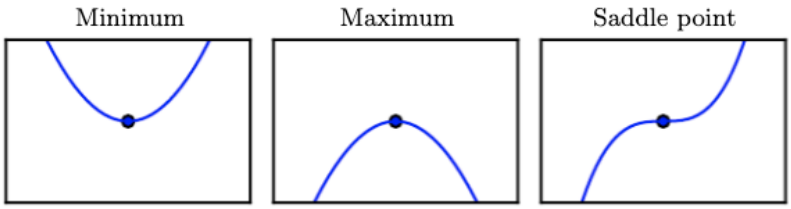

Figure 4.2: Types of critical points. Examples of the three types of critical points in one dimension. A critical point is a point with zero slope. Such a point can either be a local minimum, which is lower than the neighboring points; a local maximum, which is higher than the neighboring points; or a saddle point, which has neighbors that are both higher and lower than the point itself.

▼ 4.4

▼ 4.5

▼ Chapter 5 - Linear Algebra