

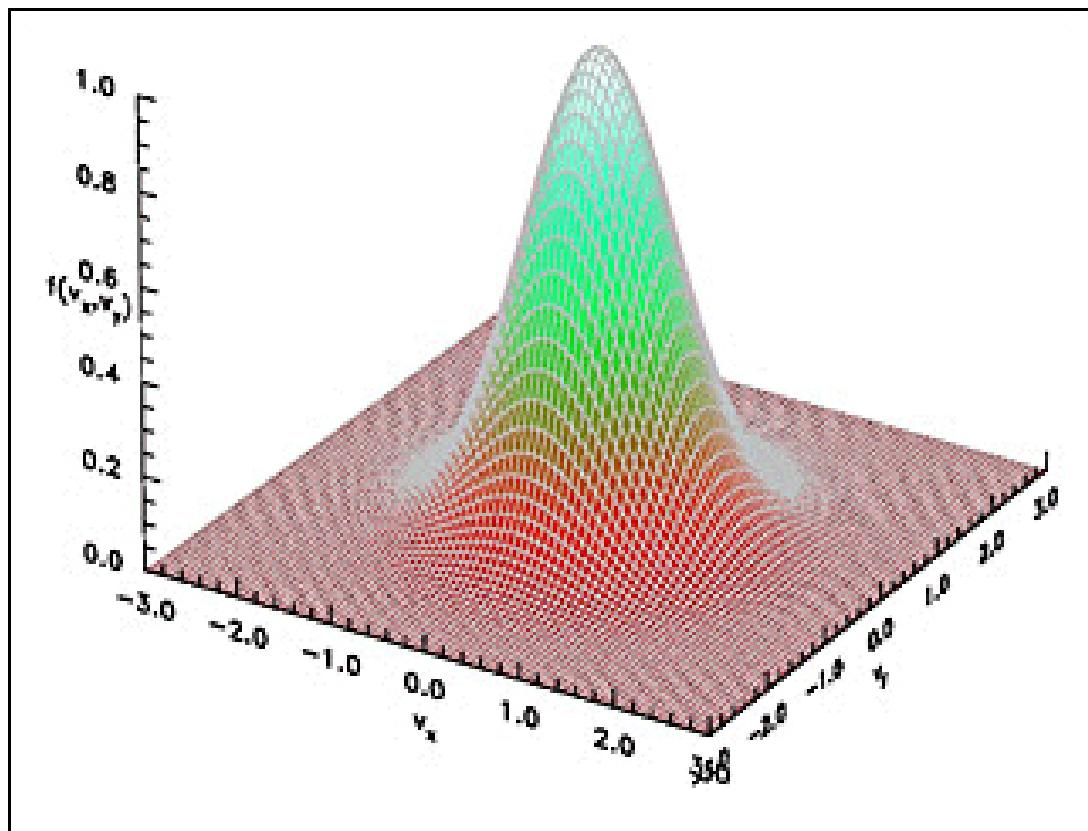
# CSIE5123: Numerical Method

Vinsong

January 16, 2026

## Abstract

The lecture note of 2023 Spring Numerical Method by professor 林智仁.



# Contents

# Chapter 1

## Floating-point systems

### Lecture 1

#### 1.1 Floating-point basics

This chapter is mainly based on the science of floating-point arithmetics which is based on the IEEE standard 754.

##### 1.1.1 Why learning floating-point operations?

**Example.** A one-variable problem

$$\min_x f(x) \quad \text{where } x \geq 0$$

In the normal program, we should set an **upper bound** of  $x$ , or  $x$  may be wrongly increased to  $\infty$ . We have to find the largest representable number in the computer

**Example.** A ten-variable problem

$$\min_{\mathbf{x}} f(\mathbf{x}) \quad \text{where } x_i \geq 0, i = 1, 2, \dots, 10$$

We want to know how many are zeros, we may use

```
1 for (int i = 0; i < 10; i++)
2     if (x[i] == 0) count++;
```

But people say that don't do the comparison of floating-point

```
1 double epsilon = 1.0e-12;
2 for (int i = 0; i < 10; i++)
3     if (x[i] <= epsilon) count++;
```

Which is better? How to chose `epsilon`? Can't do the comparison of floating-point? We need to understand the floating-point representation.

##### 1.1.2 Floating-point Format

We know `float` (single precision): 4 bytes and `double` (double precision) 8 bytes in C/C++.

**Definition 1.1.1 (format).** A floating-point system requires a base  $\beta$ , precision  $p$ , significand (mantissa)  $d.d\dots d$