南京大学

本 科 毕 业 论 文

院　　系　物理学院

专　　业　光电科学系

论 文 题 目　一种无导数优化方法及其应用

年　　级　**2012** 级　学　　号　**121120173**

学 生 姓 名　张楚珩

指 导 老 师　宋凤麒　职　　称　教授

论 文 提 交 日 期　**2016** 年 **5** 月 **20** 日

A Type of Derivative-free Optimization Approach and Its Applications

by

ZHANG Chuheng

Supervised by

Professor SONG Fengqi

An Undergraduate Dissertation Submitted to

Department of Physics, Nanjing University

Department of Physics

Nanjing University

May 20, 2016

南京大学本科生毕业论文(设计、作品)中文摘要

题　　目：一种无导数优化方法及其应用

院系：物理学院 专业：光电科学系 年级：2012 级 姓名：张楚珩

指导老师（姓名、职称）：宋凤麒教授

摘　　要：

优化理论和方法解决的是如何在给定的目标函数下选取最优决策的问题，由于它在众多领域的广泛应用，近年来已经成为了一个热门的研究领域。无导数优化方法是优化方法中的一个重要组成部分。在实际问题中，一方面由于梯度的获取有时非常困难，另一方面由于梯度方法容易陷入局部极小，因而无导数方法的重要性得以凸显。本文介绍了一种最新的无导数优化算法——Racos，它基于分类模型框架，在高维搜索问题和复杂优化问题上具有良好的性能。

本文给出了此无导数优化方法在物理器件设计和神经网络训练上的实际应用，用以说明无导数优化方法两方面的作用：在梯度无法获取情形下优化方法的直接应用；在梯度可以获取情形下结合无导数优化方法能起到更好的效果。

1）本文给出了此无导数方法在光学器件设计中的应用。本文将无导数优化方法应用到了超透镜的设计中，给出了一种消色差聚焦超透镜的设计方案。本文还将无导数优化方法应用到了硅基波导器件的设计上，我们给出了两种不同功能的器件（换模器和偏振分束器）的设计方案，它们都比现有传统硅基波导器件具有更小的特征尺度和更好的性能。这一系列的设计将有望在近期得以加工并测试。

2）本文给出了此无导数优化方法在人工神经网络中的应用。神经网络由于在处理复杂问题时表现出的强大归纳能力，近年来得到了深入的研究和广泛的应用。而神经网络的训练作为神经网络研究的一个核心问题，符合无约束优化的基本框架，是一个典型的优化问题。基于梯度的算法已被广泛而有效地应用于神经网络的训练中，当初始值在全局极小附近时，它具有良好的收敛性能，然而却容易陷入局部极小。本文尝试将无导数优化方法和基于梯度的方法相结合，得到了一些混合策略，以期望能克服纯梯度算法的一些缺陷，并且在实验上取得了比原有算法更好的结果。

关键词：　无导数优化方法；神经网络；消色差超透镜；换模器；偏振分束器

南京大学本科生毕业论文 (设计、作品) 英文摘要

THESIS: A Type of Derivative-free Optimization Approach and Its Applications

DEPARTMENT: Department of Physics

UNDERGRADUATE: ZHANG Chuheng

MENTOR: Professor SONG Fengqi

ABSTRACT:

Optimization theory or optimization method is aimed at finding and adopting a *best* strategy given a specified goal or objective. It becomes quite an active research field due to its broad applications in many fields such as mechanics, economics, electrical engineering, petroleum engineering, molecular modelling and transportation network. Derivative-free optimization method is an important category of optimization method. The importance of derivative-free optimization is significant, not only for the difficulty in obtaining the derivative of objective functions in many practical problems but also for gradient based method's suffering from innate deficiency of getting trapped in local minimum. In this dissertation, a novel type of derivative-free optimization algorithm, *Racos*, is introduced. It shows good performance in complex optimization tasks and great extensibility in high dimension optimization problems.

The latter half of the dissertation gives some specific applications of the above mentioned optimization algorithm, for the demonstration of the two roles of this optimization algorithm - the direct application of optimization method in cases where derivative is hard to obtain, and the performance promotion when combined with derivative-based method in cases where derivative is available.

1) Application of the above mentioned optimization algorithm in the field of optical device designing is demonstrated. Derivative-free optimization is applied to the designing of metasurface lens which results in a kind of design of achromatic metasurface lens. Derivative-free optimization is also applied to the designing of silicon based waveguide device, resulting in the design of two type of functional devices - mode converter and polarizing beamsplitter. Compared with traditional silicon based waveguide device, they have smaller footprints but with better performance. The above designs will be followed by further fabrication and test shortly.

2) Application of the above mentioned optimization algorithm in the field of artificial neural network is presented. In recent years, due to its great generalization ability when encountering complex problems, artificial neural network is intensely studied and

applied to many practical problems. Furthermore, the training of neural network, as a central task in the field of artificial neural network, fits the framework of unconstrained optimization, and therefore becomes a typical optimization problem. Gradient based method is widely and efficiently applied to the training of neural network. It demonstrates good performance near minimum but easily gets trapped in local minimum and falls to find a global optimal result. We attempt to combine the derivative-free algorithm with present gradient based algorithms in order to overcome the deficiencies of gradient based algorithms. Several of our hybrid strategies demonstrate better performance than the original gradient based algorithm experimentally.

KEY WORDS: Derivative-free Optimization, Neural Network, Achromatized Metasurface Lens, Mode Converter, Polarizing Beamsplitter

目　　录

目　　录 ····· **V**

插图清单 ····· **VII**

1　优化方法简介 ····· **1**

1.1　优化问题 ····· 1

1.2　基于导数的优化方法 ····· 4

1.2.1　最速下降法 ····· 4

1.2.2　牛顿法 ····· 5

1.3　无导数优化方法 ····· 6

1.3.1　坐标搜索 ····· 8

1.3.2　遗传算法 ····· 9

2　一种无导数优化算法 ····· **11**

2.1　分类算法框架 ····· 11

2.2　Racos 算法 ····· 12

3　Racos 在光学器件设计中的应用 ····· **17**

3.1　Racos 在消色差超透镜设计中的应用 ····· 18

3.1.1　超透镜简介 ····· 18

3.1.2　消色差超透镜 ····· 19

3.1.3　消色差超透镜的正向设计 ····· 22

3.1.4　消色差超透镜的逆向设计 ····· 22

3.1.5　消色差超透镜的设计结果 ····· 23

3.2　Racos 在微纳光子学器件设计中的应用 ····· 25

3.2.1　微纳光子学器件设计的物理基础 ····· 26

3.2.2　换模器 ····· 26

3.2.3 偏振分束器 ····· 27

4 Racos 在神经网络中的应用 ····· 33

4.1 神经网络简介 ····· 33

4.2 基于梯度的神经网络训练方法 ····· 35

4.2.1 神经网络的数学表达 ····· 35

4.2.2 神经网络的训练数据集 ····· 36

4.2.3 神经网络在数据集上的误差 ····· 36

4.2.4 神经网络的 BP 算法 ····· 36

4.3 结合无导数优化方法的混合策略 ····· 37

4.4 实验结果 ····· 40

5 结论 ····· 45

参考文献 ····· 47

致　　谢 ····· 53

插图清单

1-1 **基于导数的优化方法通常容易陷入局部极小值点** ····· 7

2-1 **Racos 算法在六驼峰函数上的优化。**蓝色圆点为每一代的采样点，红色十字为被标记出的“好”采样点，绿色方框为分类算法给出来的假设。每一轮迭代中，Racos 对种群进行采样并标记，并根据标记结果给出假设，使用假设来确定下一轮的采样点。····· 15

3-1 **传统透镜受到色差的影响。**由于不同波长相对于介质的折射率不一样，导致不同波长的光经过透镜时的偏折不一样，从而产生色差。····· 19

3-2 **消色差超透镜的基本原理。**(a) 如果超透镜对于任何波长的光都附加相同的相位，不同补偿的光将无法聚焦到同一点，产生色散。(b) 按照一定的规律设计超透镜各个位置对于不同波长的相位补偿可以消除色散。 ····· 20

3-3 **超透镜结构。**超透镜由介质衬底和上面的金属小棒阵列构成。每个金属小棒单元拥有不同的转角，不同数目和大小的小棒，以及不等的小棒间间距。····· 20

3-4 φ_{m2} **相对于波长的关系。**在所研究的波长范围内，附加相位相对于波长近线性变化。参数 λ_1、λ_2 和 φ_{m2} 可以确定此消色差单元。 · 21

3-5 **超透镜消色差优化结果。**三行分别表示 1）无消色差设计下超透镜的仿真结果；2）正向设计下的消色差超透镜仿真结果；3）逆向设计下的消色差超透镜仿真结果。三列分别表示 1）沿中轴线的场强分布；2）消色差部分的设计；3）转角部分的设计。····· 24

3-6 **消色差超透镜的模拟结果。**(a) 超透镜柱镜设计方案。(b)3000nm ~ 4286nm 波段内聚焦涨落在 ±5.9% 以内，红色虚线表示各波长聚焦位置，黑色虚线表示平均聚焦位置。····· 25

3-7 **换模器设计方案与模拟结果。**(a) 换模器的设计方案，其设计区域为长 $1.33\mu m$，宽 $2.81\mu m$ 的区域，每个像素尺寸 $37nm \times 37nm$，工作波长为 $\lambda = 1550nm$。(b) 换模器的模拟结果。 …… 28
3-8 **消除偏振敏感性的实验方案。**它将耦入的光按照不同偏振方向分开，分别进行处理后再合起来。 …… 29
3-9 **偏振分束器几何结构和尺寸。**波导横截面宽 $440nm$，高 $300nm$，出射波导管间距 $1\mu m$，偏振分束器大小 $2.4\mu m \times 2.4\mu m$，每个像素大小 $120nm \times 120nm$，偏振分束器工作波长为 $1550nm$。 …… 30
3-10 **偏振分束器设计方案与模拟结果。**(a) 偏振分束器的设计方案；(b) 入射波导模式为 TM 模式时的电场分布；(c) 入射波导模式为 TE 模式时的电场分布。 …… 31

4-1 **BP 神经网络的结构。**它包括输入层、隐含层和输出层；每一层都含有数目不同的神经元，每一个神经元都与相邻层的所有神经元相连。 …… 34
4-2 **BP 神经网络中神经元的结构。**前一层的信号分别于相应权值相乘后累加，再通过一个非线性映射得到该神经元的输出。 …… 35
4-3 **MNIST 数据集。**它的样本输入为大量手写数字的图片，目标输出为每张图片所对应的数字。 …… 41
4-4 ***Racos(BP)*、*Racos*×*BP* 和 *BP* 在前 50 单位计算成本下神经网络误差率的下降趋势。**x 轴表示计算成本，y 轴表示神经网络在 MNIST 测试数据集上的误差率。在较小的计算成本下，*Racos*×*BP* 能快速找到比 *BP* 更好的解。小叉表示 60 次平行测试的平均值，误差棒表示其标准差。 …… 42
4-5 **各个算法在固定 30000 单位计算成本下的误差率与 η 的关系。**x 轴表示比例 η，y 轴表示 *Racos*+*BP*、*Racos(BP)*+*BP* 和 *Racos*×*BP*+*BP* 在固定 30000 单位计算成本下的误差率。从 $\eta = 0$ 和 $\eta = 1$ 端点可以看出 *Racos*、*Racos(BP)*、*Racos*×*BP* 和 *BP* 的误差率。圆圈表示 60 次平行测试的平均值，误差棒表示其标准差。 …… 43

第一章　优化方法简介

本章主要介绍优化方法相关的一些基础知识，以保证本文的连贯性。本章介绍的一些基础优化算法与之后要介绍的算法在本质上有着密切的联系，通过对于它们的介绍，读者可以更容易地理解本文将要介绍的这种算法。

1.1　优化问题

优化问题，简言之就是在自变量定义域中选择一个较优自变量使得因变量最优。优化问题（Optimization），也有文献称最优化问题，鉴于在许多问题中难以找到最优的解，同时在很多情况下优化目标并非是找到最优的解，而是快速找到一个较优的解，本文认为称优化问题更佳。由于这种结构的问题广泛存在于包括机械制造、金融工程、电气工程、石油工程、分子模拟、交通运输在内的众多领域中，因此它也越来越受到科研机构和工业部门的重视。计算机计算性能的提高使得优化方法能够用于更多复杂的计算与模拟环境中。同时随着计算机科学的不断发展，优化计算方法也在不断进步，为优化方法提供了更多发挥的空间。

优化问题数学模型的一般形式为:

$$\begin{aligned} &arg\min_{\boldsymbol{x}} f(\boldsymbol{x}) \\ &\text{subject to } \boldsymbol{x} \in \Omega \end{aligned} \tag{1-1}$$

其中，称函数 $f:\mathbb{R}^n \to \mathbb{R}$ 为目标函数，它是一个实函数。在某些优化问题中，函数 f 的目标值可以为一个向量，这时它描述一个多目标优化问题。在这里我们仅讨论单一目标的优化问题。$\boldsymbol{x}$ 是一个 n 维的向量，称决策变量，表示为 $\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$，其中 $x_1, x_2, \cdots, x_n$ 相互独立。集合 Ω 是 n 维实数空间 $\mathbb{R}^n$ 的一个子集，称约束集。在某些问题中，还可能含有 $\boldsymbol{g}(\boldsymbol{x}) \le 0$，$\boldsymbol{g}(\boldsymbol{x}) = 0$，$\boldsymbol{A}\boldsymbol{x} \le \boldsymbol{b}$ 或者 $\boldsymbol{A}\boldsymbol{x} = \boldsymbol{b}$ 形式的约束，在多数情况下这样约束可以转化为 $\boldsymbol{x} \in \Omega$ 的形式。为了简洁起见，我们仅考虑形如 $\boldsymbol{x} \in \Omega$ 的约束条件。

在某些情况下，优化问题考虑最大化函数 f，在这些情况下我们可以通过取 $f \leftarrow -f$ 转化为最小化问题。不失一般性，我们仅讨论最小化优化问题。

优化问题的目标是找到全局极小值点，下面给出全局极小值点的定义。

定义 1-1 全局极小值点：设 n 元实值函数 $f : \mathbb{R}^n \to \mathbb{R}$，定义域为 $\Omega \in \mathbb{R}^n$，存在 $\boldsymbol{x}^* \in \Omega$，如果 $\forall \boldsymbol{x} \in \Omega \backslash \{\boldsymbol{x}^*\}$，都有

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \tag{1-2}$$

则称 $\boldsymbol{x}^*$ 为全局极小值点；当 (1-2) 式中的小于等于号取做小于号时称 $\boldsymbol{x}^*$ 为严格全局极小值点

在实际优化中，全局极小值点常常难以找到，通常情况下我们会试图寻找一个较好的局部极小值点。在数学上，局部极小值点也比全局极小值点更容易处理，事实上，我们在分析很多算法相关性质（比如收敛条件和收敛速度）的时候，常常针对局部极小值点进行分析。下面给出局部极小值点的定义。

定义 1-2 局部极小值点：设 n 元实值函数 $f : \mathbb{R}^n \to \mathbb{R}$，定义域为 $\Omega \in \mathbb{R}^n$。对于定义域中的一个点 $\boldsymbol{x}^* \in \Omega$，存在 $\epsilon > 0$，对于所有满足 $\|\boldsymbol{x} - \boldsymbol{x}^*\| < \epsilon$，$\boldsymbol{x} \in \Omega \backslash \{\boldsymbol{x}^*\}$ 的 $\boldsymbol{x}$ 都有

$$f(\boldsymbol{x}^*) \leq f(\boldsymbol{x}) \tag{1-3}$$

则称 $\boldsymbol{x}^*$ 为局部极小值点；当 (1-3) 式中的小于等于号取做小于号时称 $\boldsymbol{x}^*$ 为严格局部极小值点。

目标函数的导数不仅能判定极小值点的存在，同时对于优化问题的求解也发挥着很大的作用。下面给出多元实值函数一阶导数（梯度）和二阶导数（Hessian 矩阵）的定义。

定义 1-3 梯度：设 n 元实值函数 $f : \mathbb{R}^n \to \mathbb{R}$ 可微，$\boldsymbol{x} = [x_1, x_2, \cdots, x_n]^T$，为函

数的自变量，那么称函数

$$\nabla f(\boldsymbol{x}) = \begin{bmatrix} \dfrac{\partial f}{\partial x_1}(\boldsymbol{x}) \\ \vdots \\ \dfrac{\partial f}{\partial x_n}(\boldsymbol{x}) \end{bmatrix} \tag{1-4}$$

为 f 的梯度。

定义 1-4 Hessian 矩阵：设 n 元实值函数 $f : \mathbb{R}^n \to \mathbb{R}$，它的梯度 $\nabla f(\boldsymbol{x})$ 可微，则称函数 f 二次可微，称函数

$$H(f) = \begin{bmatrix} \dfrac{\partial^2 f}{\partial x_1^2}(\boldsymbol{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_1 \partial x_n}(\boldsymbol{x}) \\ \vdots & \ddots & \vdots \\ \dfrac{\partial^2 f}{\partial x_n \partial x_1}(\boldsymbol{x}) & \cdots & \dfrac{\partial^2 f}{\partial x_n^2}(\boldsymbol{x}) \end{bmatrix} \tag{1-5}$$

为 f 的 Hessian 矩阵。

下面给出局部极小值点需满足的必要条件。

定理 1-1 局部极小值点的一阶必要条件：设 n 元实值函数 $f : \mathbb{R}^n \to \mathbb{R}$ 在约束集 Ω 上一阶连续可微，即 $f \in C^1$，定义域 $\Omega \in \mathbb{R}^n$。如果 $\boldsymbol{x}^*$ 是函数 f 在 Ω 上的局部极小值点，并且是 Ω 的内点，则有

$$\nabla f(\boldsymbol{x}^*) = 0 \tag{1-6}$$

定理 1-2 局部极小值点的二阶必要条件：设 n 元实值函数 $f : \mathbb{R}^n \to \mathbb{R}$ 在约束集 Ω 上二阶连续可微，即 $f \in C^2$，定义域 $\Omega \in \mathbb{R}^n$。如果 $\boldsymbol{x}^*$ 是函数 f 在 Ω 上的局部极小值点，并且是 Ω 的内点，则有

$$\nabla f(\boldsymbol{x}^*) = 0 \tag{1-7}$$

并且，Hessian 矩阵 $\boldsymbol{H}(f)$ 为半正定矩阵，即 $\forall \boldsymbol{d} \in \mathbb{R}^n$，有

$$\boldsymbol{d}^T \boldsymbol{H}(f) \boldsymbol{d} \leq 0 \tag{1-8}$$

限于篇幅仅讨论极值点是定义域内点的情形，极值点位于定义域边界上的情形请参见文献[1]。

1.2 基于导数的优化方法

目标函数的导数在优化问题中包含了大量的有用信息。正如我们前面所看到的，优化问题的主要目的是要找到目标函数的极小值点，而极小值点的必要条件也正由目标函数的导数决定，即目标函数的梯度和 Hessian 矩阵需满足 (1-7) 式和 (1-8) 式的关系。因此，在导数容易求得的情况下，有效利用导数信息能够有效帮助我们快速找到优化问题的解。

基于导数的优化方法主要有最速下降法、牛顿法、共轭梯度法以及牛顿法的一系列衍生算法。我们这里主要简要介绍最速下降法和牛顿法，以说明一阶导数和二阶导数在优化方法中所起到的作用。我们希望通过列举这两种典型的算法说明当导数信息可用的时候我们如何是如何利用它们的，以及导数信息对于带给优化问题的便利。

1.2.1 最速下降法

考虑目标函数 $f:\mathbb{R}^n\to\mathbb{R}$ 和决策变量 $\boldsymbol{x}$，目标函数 f 点 $\boldsymbol{x}$ 处在 $\boldsymbol{d}$ 方向上增长率可写为 $\langle\nabla f(\boldsymbol{x}),\boldsymbol{d}\rangle$，$\|\boldsymbol{d}\|=1$，由 Cauchy–Schwarz 不等式可知

$$\langle\nabla f(\boldsymbol{x}),\boldsymbol{d}\rangle\le\|\nabla f(\boldsymbol{x})\| \tag{1-9}$$

并且当且仅当 $\boldsymbol{d}=\nabla f(\boldsymbol{x})/\|\nabla f(\boldsymbol{x})\|$ 的时候，增长率最大。即梯度方向是函数 f 增长最快的方向，类似地，梯度的负方向 $-\nabla f(\boldsymbol{x})$ 是函数值减小最快的方向。由此可知，如果需要搜索函数的极小值点，梯度负方向是一个很好的搜索方向。

基于此我们得到最速下降法的算法。

注意到，α_k 的选择中还包括了一个一维的优化问题，可以利用多种一维搜索方法求得，这里不再赘述。当 $\|\nabla f(\boldsymbol{x}^*)\|=0$ 的时候，满足局部极小值点的一阶必要条件，在很多情况下，这就是我们要找到的局部极小值点，因此

算法 1.1 最速下降法

1: **repeat**
2: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})$
3: where $\alpha_k = arg\min_{\alpha \geq 0} f(\boldsymbol{x}^{(k)} - \alpha \nabla f(\boldsymbol{x}^{(k)}))$
4: **until** $\|\nabla f(\boldsymbol{x})\| < \epsilon$

可以用 $\|\nabla f(\boldsymbol{x})\| < \epsilon$ 作为一个停止规则。当然还有许多其他的停止规则，比如 $\|\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}\| < \epsilon$，$\|f(\boldsymbol{x}^{(k+1)}) - f(\boldsymbol{x}^{(k)})\| < \epsilon$ 等。

最速下降法具有良好了理论基础，它被证明对于任意的起点，都能产生一系列的迭代点序列，最终收敛到满足局部极小值点一阶必要条件的点（全局收敛性）和至少 1 阶的收敛阶数。[1]

值得一提的是形如 $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \alpha_k \nabla f(\boldsymbol{x}^{(k)})$ 的迭代方式是许多基于梯度方法的一个基本思路，在本文之后所介绍的神经网络训练的 BP 算法也用到了这样的迭代形式。

1.2.2 牛顿法

最速下降法中利用到了目标函数的一阶导数，有些时候目标函数的二阶导数也可以轻易获得，在优化过程中使用二阶导数的性质可以使得优化更为高效。牛顿法就是同时利用了一阶和二阶导数的信息，在初始点与目标函数极小值点足够接近的时候，牛顿法具有更高的效率。牛顿法算法概括如下。

算法 1.2 牛顿法

1: **repeat**
2: $\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \boldsymbol{H}(\boldsymbol{x}^{(k)})^{-1} \nabla f(\boldsymbol{x}^{(k)})$
3: **until** $\|\nabla f(\boldsymbol{x})\| < \epsilon$

其中，$\boldsymbol{H}(\boldsymbol{x}^{(k)})$ 是目标函数的 Hessian 矩阵。牛顿法的实质是在给定一个迭代点之后，构建一个在迭代点处一阶导数和二阶导数都与迭代点相同的二次型函数，然后将估计的二次型函数的最小值点作为下一个迭代点。

虽然牛顿法收敛比较快，但是它具有以下两方面的缺点：1）牛顿法的收敛条件比较苛刻，存在牛顿法失效的情形（比如目标函数 Hessian 矩阵非正定）；2）计算量较大：牛顿法需要计算 Hessian 矩阵，在决策变量维度较高的时候，Hessian 矩阵的规模随着维度 n 以 n^2 增长，同时对于 Hessian 矩阵的求逆

也需要较大的计算量。通过 Levenberg-Marquardt 修正和包括 DFP 和 BFGS 在内的拟牛顿方法可以使得以上缺陷得以弥补。

1.3 无导数优化方法

尽管对于导数信息的利用能使我们更快找到优化问题的局部极小值解，无导数优化方法在很多地方也发挥着其不可替代的作用。无导数优化方法主要应用于两种情形：1）导数的获取非常困难的情形；2）导数常常误导优化过程陷入离初始点较近的局部极小，而某些时候更需要一个离初始点更远而更优的解。

在许多现实的问题中，导数常常不可直接获得。鉴于导数信息在优化问题中所发挥的巨大作用，在导数不可直接获得的某些情况下，仍然有一些方法可以使得我们获得相应的导数。我们可以 1）使用一些自动求导软件进行求导，或者 2）利用有限差分方法求得近似的导数值。

然而，在实际应用中，仍然存在这两种方法失效的情形。

对于第一种方法而言，它需要要求目标函数能够解析地表达出来。在许多应用场合下，目标函数值的求取是通过封装好的程序或者函数求得的，因而我们无法获知目标函数的解析表达式。同时，在包含物理器件设计、分子模拟、飞行器设计、石油勘探等许多需要优化设计的领域，目标函数值的获取通常包含许多模块的顺序执行，比如几何结构的生成、网格的划分、偏微分方程的求解和相应的后处理模块，这使得目标函数相对于决策变量的关系变得非常复杂，以至于无法求得其相应的导数。即使在某些我们能写出目标函数解析表达式的情况下，它的导数也可能不存在或者无法求出。

对于第二种方法而言，它在每一次迭代中都需要大量查询求取目标函数的值。举例来说，对于决策变量为 n 维的问题，如果需要使用有限差分方法求其梯度，则需要计算 n 次目标函数的值；如果需要求其 Hessian 矩阵，则需要计算 $n(n-1)/2$ 次目标函数的值。在目标函数值的求取代价非常高昂的情况下，这显然是不太合适的。

可见，在许多导数获取非常困难的情况下，无导数优化方法具有不可替代的作用。

在复杂优化问题中，全局最优解常常是难以找到的。同时目前也没有有效的算法和理论能保证找到全局最优解，除非目标函数具有某些良好的性质，比如目标函数上所有的局部极小值点都是全局极小值点，例如凸函数。基于导数的算法通常依赖于当前点的导数信息，从而仅仅找到离初始点较近的局部极小值点。如图 1-1所示，基于导数的算法通常会陷入一个不是最优的极小值点。而通过一些无导数的优化算法，我们可以找到一个全局上相对更优的解。值得注意的是，由于这些全局无导数算法中应用到的启发式搜索过程，它们的理论基础都较为薄弱，但是它们在实践中通常都拥有比基于导数的算法更好的效果。

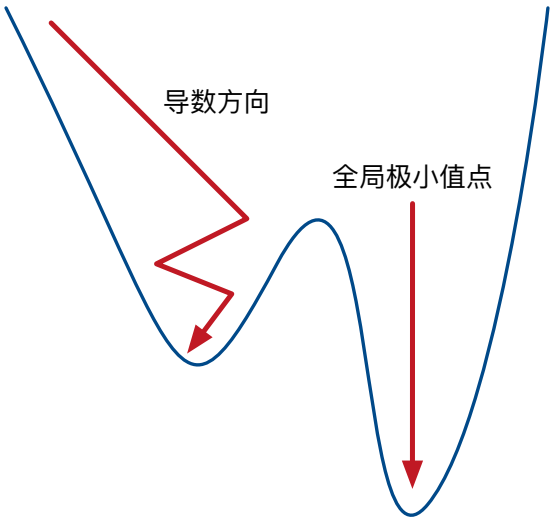

图 1-1: **基于导数的优化方法通常容易陷入局部极小值点**

可见，如果函数的形貌较为复杂的时候，也常常需要无导数方法进行优化。

综上所述，尽管导数信息在优化中发挥着很大的作用，无导数方法仍然在很多地方有着不可替代在作用。它已经被越来越多地用于工业设计和其他学科的科研工作当中，有意思的应用包括在分子模拟中的应用[2]、在飞行器设计中的应用[3,4]、在流体力学中的应用[5,6]、在医学中的应用[7,8] 和在地质学中的应用[9–12] 等。

局部无导数优化方法主要有直接搜索的方法（Direct Local Search）[13]、和基于模型的方法（Local Model-based Method）。具体说来直接搜索的方法有 Nelder-Mead 单纯形算法（Nelder–Mead simplex algorithm）[14]、GPS（Gen-

eralized Pattern Search）[15]、GSS（Generating Set Search）[16] 等。基于模型的方法主要有信赖域方法（Trust-region methods）[17]、隐性过滤（Implicit Filtering）[18] 等。

全局无导数优化方法主要包括各种演化算法（Evolutionary Algorithm），比如遗传算法（Genetic Algorithm）[19]、粒子群算法（Particle Swarm Algorithm）[20]、模拟退火算法（Simulated Annealing）[21]、分布估计算法（Estimation of Distribution Algorithm）[22]、CMA-ES（Covariance Matrix Adaptation Evolution Strategy）[23]、REMBO（Bayesian Optimization with Random Embedding）[24]、SOO（Simultaneous Optimistic Optimization）[25] 等。

为了说明无导数方法寻找极小值点的过程，我们介绍一种最简单的坐标搜索方法。为了说明全局无导数方法所起到的作用，我们在介绍一种基本的遗传算法。同时，这两种算法的思想和框架也会用于我们后面将要着重介绍的一种无导数优化算法之中。

1.3.1 坐标搜索

下面介绍一种坐标搜索算法（Coordinate Search），它是模式搜索（Pattern Search）的原型。一个无梯度的搜索方法要收敛到极小值点，至少需要 1）采样点的均值需要保证某种形式的下降；2）控制采样点分布的形状；3）搜索步长能收敛到零。算法框架如算法 1.3 所示。坐标搜索算法在遇到一个数值更小的采样点的时候使用新的采样点来作为下一轮的采样中心（算法第 5 行），从而保证了采样点均值的总体下降趋势。它各个采样点的分布始终保持沿着坐标轴以步长为间隔分布（算法第 4 行），即搜索方向与坐标轴平行，这也是称之为坐标搜索的原因。在当前采样点比周围采样点都好的时候，说明极小值点就在当前采样点附近，这时减半步长（算法第 10 行）。通过反复的迭代收敛到极小值点处。

这是一种局部的无导数优化方法，当目标函数较为简单的时候，它能有效地找到极小值，但是当目标函数有众多极小值点的时候，其给出的解的优劣很大程度上取决于初始点的选取。

算法 1.3 坐标搜索

1: $\boldsymbol{x}_0 \leftarrow$ initial point in $\mathbb{R}^n$
2: $\Delta_0 > 0 \leftarrow$ initial mesh size
3: **for** $k = 1$ **to** k_{max} **do**
4: **if** $f(t) < f(\boldsymbol{x}_k)$ for some $t \in P_k = \{\boldsymbol{x}_k \pm \Delta_k \boldsymbol{e}_i | i = 1, \cdots, n\}$ **then**
5: $\boldsymbol{x}_{k+1} = t$
6: $\Delta_{k+1} = \Delta_k$
7: **else**
8: otherwise $\boldsymbol{x}_k$ is a local minimum with respect to P_k
9: $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k$
10: $\Delta_{k+1} = \frac{\Delta_k}{2}$
11: **end if**
12: **end for**

1.3.2 遗传算法

遗传算法的主要思想是基于生物种群的进化规律，通过遗传、变异、选择、杂交等手段，保证随着迭代的进行，采样点集合逐步更新更优的解。遗传算法的框架如算法 1.4 所示。最开始在可行域中等概率随机采样，并且计算相应的目标函数值，用以构成最初的种群（算法第 1 行和第 2 行）。在每一代中，依次生成 m 个新的个体，用以构成下一代种群。生成个体的方法是从上一代的种群中通过选择、杂交来得到（算法第 5 行）然后在进行一定概率的变异（算法第 6 行）。由于种群选择和杂交的方式在不同的遗传算法中采取不同的策略，比如轮盘法、竞争法等，这里都简记写在算法第 5 行中。

算法 1.4 遗传算法

1: collect $S_0 = \{\boldsymbol{x}_1, \cdots, x_m\}$ by i.i.d. sampling from Ω
2: evaluate $f(\boldsymbol{x})$ and construct $B_0 = \{(\boldsymbol{x}_1, y_1), \cdots, (\boldsymbol{x}_m, y_m)\}$
3: **for** $k = 1$ to k_{max} **do**
4: **for** $i = 1$ to m **do**
5: generate $\boldsymbol{x}_i$ based on B_{k-1} through recombination
6: mutate $\boldsymbol{x}_i$
7: add $(\boldsymbol{x}_i, f(\boldsymbol{x}_i))$ to B_k
8: **end for**
9: **end for**

通过上述遗传算法的框架我们可以看出这一类全局无导数算法的基本要素：1）利用一个种群在可行域上进行搜索；2）种群迭代进行更新；3）每一

代种群的采样依赖于在上一代种群中的观测到有关于目标函数的信息；4）每一代种群在总体上优于之前的总群，并最后搜索到一个较好的结果。

我们接下来要介绍的一种新的全局无导数优化算法将会利用这种演化算法的思想。

第二章　一种无导数优化算法

基于分类模型的坐标搜索算法 Racos（Radomized Coordinate Shrinking Algorithm）是由 Yang Yu 等[26] 提出的一种全局无导数优化算法。它在能够处理较为复杂目标函数的最优化问题，同时能够应对较高维度的最优化问题。在理论上，能够证明它具有良好的收敛性质，它能够在多项式时间内收敛到具有局部 Lipschitz 连续性的极小值点[26]。同时，这种方法可以同时应用于连续域和离散域的优化问题中。

这个算法是基于分类算法框架的，我们接下来先介绍分类算法框架，再介绍 Racos 具体的算法。

2.1　分类算法框架

分类算法的框架和诸多演化算法相似，它利用一个种群在可行域上观察目标函数，并通过观察到的目标函数值调整下一代的采样位置，如此循环以找到更好的解。分类算法的核心思想是通过对于采样点的采样，将采样点分为“好的”和“坏的”两类。通过对于采样点的分类，将解的可行域分为“好的”和“坏的”，在下一步的采样中，就着重在“好的”可行域中采样。如此循环，以找到较好的解。

这里考虑如式 (1-1) 描述的优化问题。分类算法的框架如算法 2.1所示。分类算法最初的种群从一个在可行域上的均匀采样开始（第 1 行），然后进行 T 次循环。其中，$\mathcal{U}_\Omega$ 表示在可行域 Ω 上的均匀采样。在每一次循环中，都会先构建一个二元分类器 $B_t = \{(\boldsymbol{x}, y)\}$（第 4 行）。其中 $\boldsymbol{x}$ 是上一代种群中的个体，y 按照如下方式定义，1）如果相应的 $\boldsymbol{x}$ 是当前种群中最好的 $\lfloor \beta \rfloor$ 个解之一，则 $y = 1$；2）否则，$y = -1$。即，$y_i = sign[\beta - rank_{S_{t-1}}[f(\boldsymbol{x}_i)]]$，其中函数 $rank_{S_{t-1}}[f(x_i)] = n$ 表示的含义是，在所有的 $f(\boldsymbol{x})$，$\forall \boldsymbol{x} \in S_{t-1}$ 中，$f(\boldsymbol{x}_i)$ 是第 n 小的。之后，通过采样过程（第 9 行），m 个个体依次生成并且加入到下一代的种群中。如此循环，当达到最大迭代数的的时候，返回当前种群中最好的结

果作为优化结果。

上述采样过程需要用到通过特定分类算法 C、基于二元分类器 B_t 得到的假设 h_t。这里提到的“假设”是一个将可行域 Ω 映射到 $\{+1,-1\}$ 的函数，我们记 $D_h = \{x \in X | h(x) = +1\}$。分类算法 C 的作用是给出一个假设 h_t 使得 $h_t(x) = +1, \forall (x,+1) \in B_t$，并且 $h_t(x) = -1, \forall (x,-1) \in B_t$。采样过程还需要用到一个平衡系数 $\lambda \in [0,1]$，它表明采样过程以 λ 的概率从 D_{h_t} 中均匀采样，以剩下的 $1-\lambda$ 的概率从整个可行域 Ω 中均匀采样。较大的 λ 会产生较快的收敛速度，但是也容易提前收敛到较差的极小值点。

算法 2.1 分类模型算法框架

1: Collect $S_0 = \{x_1, \cdots, x_m\}$ by i.i.d. sampling from $\mathcal{U}_\Omega$;
2: let $\tilde{x} = arg\min_{x \in S_0} f(x)$;
3: **for** $t = 1$ to T **do**
4: Construct $B_t = \{(x_1, y_1), \cdots, (x_m, y_m)\}$,
5: where $x_i \in S_{t-1}$ and $y_i = sign[\beta - rank_{S_{t-1}}[f(x_i)]]$
6: let $S_t = \emptyset$
7: **for** $i = 1$ to m **do**
8: $h_i = C(B_t)$, where $h_t \in \mathcal{H}$
9: $x_i = Sampling(h_t, \lambda)$, and let $S_t = S_t \cup \{x_i\}$
10: **end for**
11: $\tilde{x} = arg\min_{x \in S_t \cup \{\tilde{x}\}} f(x)$
12: **end for**
13: **return** $\tilde{x}$ and $f(\tilde{x})$

2.2 Racos 算法

不难看出，能满足前述分类算法 C 要求的算法并不唯一，这里 Racos 就提出了一种可行的分类算法。在这种算法中，D_h 一直保持超立方体的形状（各个边与坐标轴平行），并且收缩以满足 D_h 包含所有“好”的采样点，而不包括“坏”的采样点。

Racos 算法中的分类算法如算法 2.2所示。在这个分类算法的每次运行中，当前种群构成的二元分类器 B_t 都传入这个分类算法中作为输入。我们定义 B_t^+ 为 B_t 中“好”的样本集，定义 B_t^- 为 B_t 中“坏”的样本集。最开始，从 B_t^+ 中随机选一个好的样本（第 3 行），并且设假设 h_t 所划定的“好”的区域为整个

可行域 Ω（第 4 行）。在接下来的迭代过程中，坏样本被一个个选中，然后通过可行域的收缩来把该选中的坏样本排除在外（第 5-22 行）。注意到，这里针对连续和离散域上的优化都给出了相应的收缩策略。对于离散优化而言，就是把选定位元上的取值规定为当前选定的好样本在该位点的值。对于连续优化而言，就是在选定位点上，在选定的好样本和坏样本之间随机选择一个分界值把它们隔开，并以此为界收缩。同时，不确定位元的数量 N 也会被控制。如果不确定位元的数量比预期的多，则多出来的一些位元就会被选定并且规定它们的值只能为当前选定的好样本相应位元的值（第 23-26 行）。这样做是为了快速缩小下一步的搜索范围，以便能在高纬度的搜索空间中快速找到一个合适的解。

Racos 算法就是通过将上述分类算法（算法 2.2）放入分类模型框架（算法 2.1）中得到的。

Racos 算法在六驼峰函数（Six Hump Camel Function）[27] $f(x_1, x_2) = (4 - 2.1x_1^2 + \frac{x_1^4}{3})x_1^2 + x_1x_2 + (-4 + 4x_2^2)x_2^2$ 上的优化示意图如图 2–1所示。约束集 $\Omega = [0, 1]^2$，图中蓝色圆点为每一代的采样点，红色十字标出的是被标记出的“好”采样点，绿色方框是分类算法给出来的假设，背景中为目标函数的等高线。可以看出，每一代的假设都包含所有的“好”采样点，同时排除了所有的“坏”采样点。下一代的采样点来自于上一代分类算法给出的假设。这里每一代种群个体数 $m = 5$，“好”采样点的个数 $\beta = 1$，从上一次假设中采样的概率 $\lambda = 0.9$，控制的不确定位点数目 $N = 1$。值得注意的是，为了演示的直观性，这里的优化仅为二维优化，同时一个维度被固定，因此假设给出的采样范围为一条线段。当优化维度较高时，固定一些维度有助于快速缩小搜索范围并达到较优的解。

算法 2.2 Racos

B_t^+ =the positive solutions in B_t
B_t^- =the negative solutions in B_t
Randomly select $x_+ = (x_+^{(1)}, \cdots, x_+^{(n)})$ from B_t^+
let $D_{h_t} = \Omega, I = \{1, \cdots, n\}$
while $\exists x \in B_t^- : h_t(x) == +1$ **do**
  k = randomly selected index from the index set I
  $I = I - \{k\}$
  if $\Omega = \{0,1\}^n$ **then**
    $D_{h_t} = D_{h_t} - \{x \in \Omega | x^{(k)} \neq x_+^{(k)}\}$
  end if
  if $\Omega = [0,1]^n$ **then**
    x^- = randomly selected solution from B_t^-
    $B_t = B_t - \{x^-\}$
    if $x_+^{(k)} \geq x_-^{(k)}$ **then**
      r = uniformly sampled value in $(x_-^{(k)}, x_+^{(k)})$
      $D_{h_t} = D_{h_t} - \{x \in X | x^{(k)} < r\}$
    else
      r = uniformly sampled value in $(x_+^{(k)}, x_-^{(k)})$
      $D_{h_t} = D_{h_t} - \{x \in X | x^{(k)} > r\}$
    end if
  end if
end while
while $\#I > N$ **do**
  k = randomly selected index from the index set I
  $D_{h_t} = D_{h_t} - \{x \in \Omega | x^{(k)} \neq x_+^{(k)}\}, I = I - \{k\}$
end while
return h_t

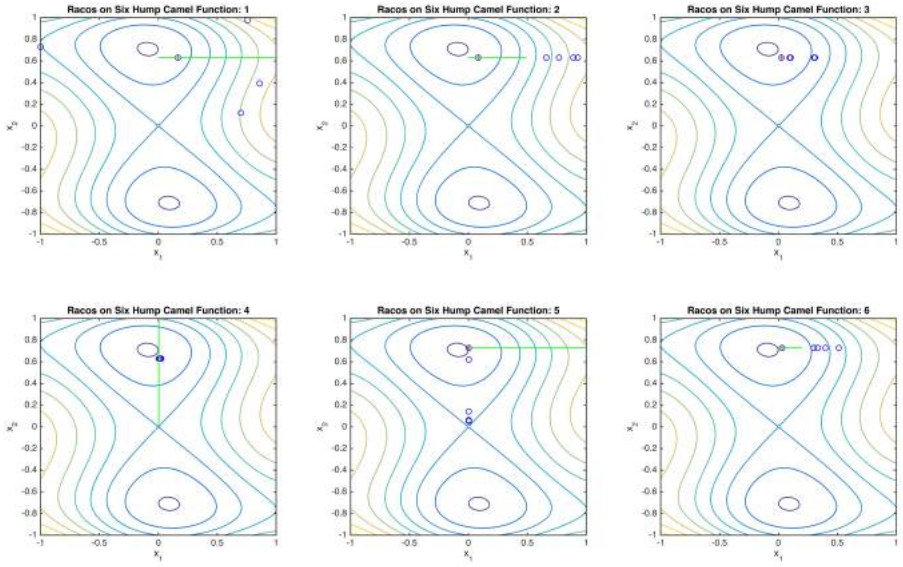

图 2-1: **Racos 算法在六驼峰函数上的优化**。蓝色圆点为每一代的采样点，红色十字为被标记出的“好”采样点，绿色方框为分类算法给出来的假设。每一轮迭代中，Racos 对种群进行采样并标记，并根据标记结果给出假设，使用假设来确定下一轮的采样点。

第三章　Racos 在光学器件设计中的应用

长期以来，光学器件的设计大多数都通过人们对于物理规律的理解，通过计算不同简单结构所产生的效应来了解结构和其功能之间的关系，通过对于许多简单结构的组合来设计满足所需功能的器件。这样的设计方法我们称之为正向设计。正向设计的过程需要人为地反复调参和试错，这样的设计效率较低。同时，由于我们通常仅仅对几种简单结构的作用比较熟悉，并把它们当做基本元件来构建我们所需要的器件，这样的设计通常没有利用到微纳加工所提供的所有自由度，因而所设计器件的效率和集成度通常没有达到最佳水平。

在光学器件的设计中，我们还可以先明确所设计器件应该满足的物理场要求，同时参考现今微纳加工水平给出物理器件的调控维度，利用一些优化算法自动给出一个最优的器件设计方案，这样的方法我们称之为逆向设计。在已有的文献中，已经有报道这种逆向设计方法在微纳光子学领域的应用[28]，但所用到的优化方法仍然比较粗糙（该文献使用 Direct Binary Search 方法进行搜索）。

Racos 作为一种无导数优化方法，对于高维度复杂函数的优化具有较好的优化性能。在光学器件的设计领域，通常需要计算在某种特定的光学器件设计下电磁场的分布状况，并且判断这样的分布是否符合设计要求。由于物理器件设计和最后电磁场分布之间复杂的物理关系，这样的映射通常非常复杂，使得难以求得其导数。同时，在某些更为复杂的模拟过程中，电磁场的模拟将会用到一些封装的有限元分析软件，这样的封装更加使得复杂的映射关系变为一个“黑箱”系统，其内部的映射结构难以判知。由此可见，Racos 这一类的无导数优化方法非常适合对于光学器件的设计进行优化。

3.1 Racos 在消色差超透镜设计中的应用

3.1.1 超透镜简介

几乎所有电磁现象和器件都是电磁波和材料相互作用产生的。从这个意义上来讲，要利用电磁波实现各种功能，就需要我们精细地选取不同的材料并为其设计结构。为了最大限度地调控电磁波，我们就必须更精细地设计材料。现今，我们对于材料的结构设计已经深入到亚波长领域，在这个尺度下制造出来的具有精细结构的材料，我们称之为超构材料（Metamaterial）。超构材料已经广泛地被应用于光学隐身设计、负折射率材料制备、微纳集成光路和超分辨成像等方面。超构材料具有许多自然界中材料难以呈现的性质，因此为光学器件的设计提供了巨大的空间，同样，超构材料也可以用于光学透镜的设计中。

使用超构材料设计得到的光学透镜可以用于实现传统光学透镜的各种功能（傅里叶变换、聚焦、成像等），这里称超透镜（Metasurface）。值得注意的是，还有一些用超透材料制备的透镜也称超透镜，它们主要用于实现近场或者远场的超分辨成像，与本章提及的超透镜功能和结构均有所不同。

光学成像系统常常受到各种像差的影响，导致了成像质量的降低。常见的五种初阶像差包括球差、慧差、像散、像场弯曲和色差。传统的透镜基于电磁波在透镜中相位连续变化来塑造成像所需要的波前，因而透镜具有一定的形貌特征而不是扁平的，所以，传统的透镜会产生球差、慧差、像散和像场弯曲。而超透镜通过光经过透镜时发生的相位突变来塑造所需波前，它是一个薄而平的表面，因此不会受到上述前四种像差的影响。

传统透镜和超透镜都会受到色差的影响。如图 3-1所示，对于传统透镜来说，在正常色散的材料中，随着波长的增大，折射率随之减小，透镜对于波长较长的光的偏折就更弱，因而对于汇聚透镜而言，其焦点就更远。不同波长的光不能汇聚到同一个焦点上，从而产生色散。如图 3-2(a) 所示，对于没有消色差功能的超透镜而言，它对于入射到它不同位置 r 的光都附加一个相位，如果这个相位对于不同波长的光都一样，那么不同波长的光在透镜后方形成的波前将会不一样，从而导致不同波长的光无法聚焦到同一个点，因此会产生色散现象。

可见，色差来源于两个方面：一方面是不同波长的光通过透镜的时候产生的偏折不一样，从而导致无法聚焦在同一焦点处；另一方面由于不同波长的光如果需要聚焦在同一焦点处，需要在透镜后方形成不同的波前，如果透镜对于不同波长的光具有一样的相位响应，那么它们最终无法汇聚到同一个焦点上。对于传统透镜来讲这两者都存在；对于普通超透镜来讲只存在后者。

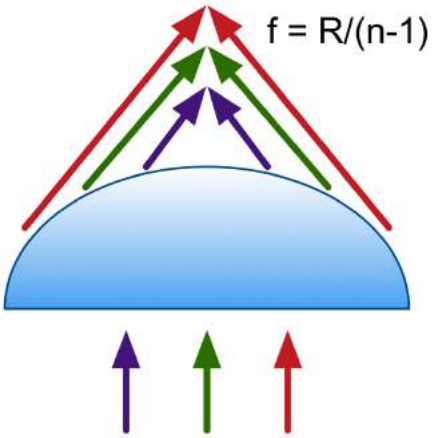

图 3-1: **传统透镜受到色差的影响。**由于不同波长相对于介质的折射率不一样，导致不同波长的光经过透镜时的偏折不一样，从而产生色差。

3.1.2 消色差超透镜

如图 3-2(b) 所示，某种波长的光经过透镜 r 处到透镜后方波前的累加相位可以表示为

$$\varphi_{tot}(r,\lambda) = \varphi_m(r,\lambda) + \varphi_p(r,\lambda) \tag{3-1}$$

其中，φ_m 是超透镜在 r 处附加的相位，φ_p 是光在自由空间中传播累积的相位，有

$$\varphi_p(r,\lambda) = \frac{2\pi}{\lambda} l(r) \tag{3-2}$$

而 $l(r)$ 与所设计光学器件所需的波前形状有关。为了能让不同波长的波都能在焦点处相干叠加，总累积相位 φ_{tot} 应该相对于波长恒定，因此要求

$$\varphi_m(r,\lambda) = -\frac{2\pi}{\lambda} l(r) \tag{3-3}$$

上式就是我们设计消色差透镜的一个基础。

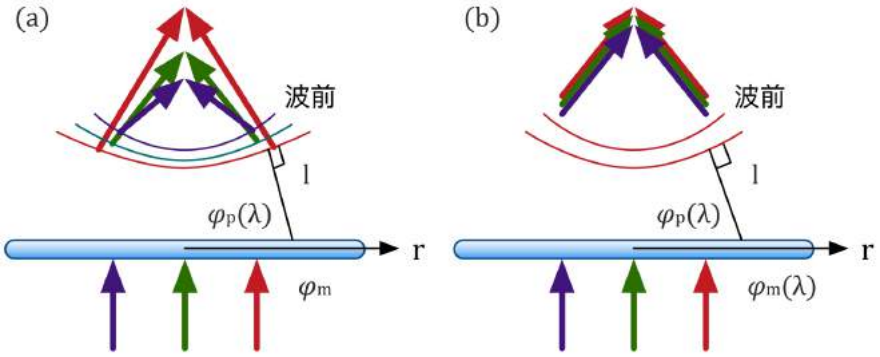

图 3-2: **消色差超透镜的基本原理。**(a) 如果超透镜对于任何波长的光都附加相同的相位，不同补偿的光将无法聚焦到同一点，产生色散。(b) 按照一定的规律设计超透镜各个位置对于不同波长的相位补偿可以消除色散。

我们采用如图3-3所示结构的超透镜。它由介质衬底和上面的金属小棒阵列构成，每个阵列单元之间的间隔 q 在 x 和 y 方向都是相同的。每个单元的小棒相对于 x 轴具有不同的转角 φ，这个转角提供了电磁波通过此单元时的一部分附加相位。同时每个单元也具有不同类型的小棒组合，不同的小棒组合提供了对于不同波长电磁波的相位补偿。

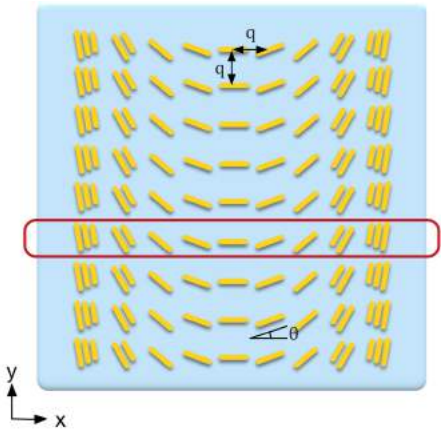

图 3-3: **超透镜结构。**超透镜由介质衬底和上面的金属小棒阵列构成。每个金属小棒单元拥有不同的转角，不同数目和大小的小棒，以及不等的小棒间间距。

每个单元中金属小棒的转角可以使得电磁波在通过此单元的时候获得一个附加相位，附加相位大小为 $\varphi_{m1} = 2\theta$。具体推导过程如下。考虑一束垂直入射到上述超透镜表面的的电磁波，超透镜表面金属小棒相对于 x 轴的夹角为 θ。

金属小棒可以看做一个电偶极子，入射光激发的电极化矢量可以写为

$$\begin{pmatrix} p_x \\ p_y \end{pmatrix} = \alpha_e \begin{pmatrix} \cos^2\varphi & \sin\varphi\cos\varphi \\ \sin\varphi\cos\varphi & \sin^2\varphi \end{pmatrix} \begin{pmatrix} E_x \\ E_y \end{pmatrix} \tag{3-4}$$

利用圆偏正光和线偏振光的基底变换关系 $\boldsymbol{e}_{L(R)} = (\boldsymbol{e}_x \pm i\boldsymbol{e}_y)/\sqrt{2}$，可以算出圆偏振光 $\boldsymbol{E}_{L(R)} = E(\boldsymbol{e}_x \pm i\boldsymbol{e}_y)/\sqrt{2}$ 入射时电极化矢量的响应

$$\begin{aligned} \boldsymbol{P}_{L(R)} &= p_x\boldsymbol{e}_x + ip_y\boldsymbol{e}_y \\ &= \frac{1}{2}\alpha_e E(\boldsymbol{e}_x \pm i\boldsymbol{e}_y) + \frac{1}{2}\alpha_e e^{\pm i2\varphi} E(\boldsymbol{e}_x \mp i\boldsymbol{e}_y) \\ &= \frac{1}{\sqrt{2}}\alpha(\boldsymbol{e}_{L(R)} \pm e^{\pm i2\varphi}\boldsymbol{e}_{R(L)}) \end{aligned} \tag{3-5}$$

可以看出，一束左旋圆偏振光入射可以得到一束附加有 2φ 相位的右旋圆偏振光，反之亦然。

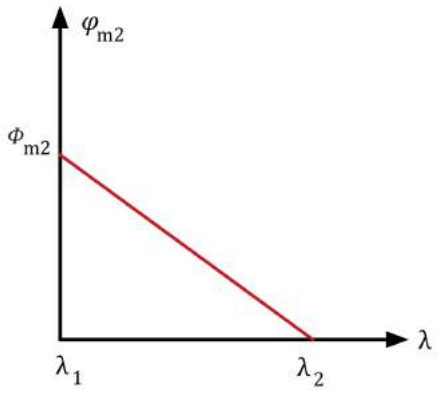

图 3-4: φ_{m2} **相对于波长的关系。**在所研究的波长范围内，附加相位相对于波长近线性变化。参数 λ_1、λ_2 和 φ_{m2} 可以确定此消色差单元。

通过调控每个单元金属棒的根数和形状可以使得不同波长的电磁波获得另外的相位补偿 φ_{m2}，它随波长近线性变化。我们设计得到了不同的结构单元，使得每个单元能够对于不同波长电磁波进行相应的相位补偿，通过这种方式的相位补偿 φ_{m2} 可以和转角附加相位 φ_{m1} 叠加，从而实现对于不同波长电磁波通过超透镜时附加相位更精细的调控。φ_{m2} 相对于波长的关系如图3-4所示。其中 λ_1 和 λ_2 是该消色差超透镜工作波长的两端值。通过消色差参数 Φ_{m2} 可以确定某个单元消色差结构的种类，在我们设计的消色差透镜中，Φ_{m2} 取 0 到 Φ_{max} 之

间均匀分布的一些列离散值，其中 $\Phi_{\max}$ 为其最大值。

通过结合以上两种调控方法，在原理上我们可以设计出满足 (3-3) 式要求的 $\varphi_m(r,\lambda)=\varphi_{m1}(r)+\varphi_{m2}(r,\lambda)$，从而实现消色差超透镜。

3.1.3 消色差超透镜的正向设计

对于聚焦焦距为 f 的超透镜而言，有 $l(r)=\sqrt{f^2+r^2}-f$。(3-3) 式可以分解为

$$\begin{aligned}|\varphi_m| &= \frac{2\pi}{\lambda_2}l(r)+2\pi l(r)(\frac{1}{\lambda}-\frac{1}{\lambda_2}) \\ &= \frac{2\pi}{\lambda_2}l(r)+2\pi l(r)(\frac{1}{\lambda_1}-\frac{1}{\lambda_2})(\frac{\lambda_2-\lambda}{\lambda_2-\lambda_1}) \\ &= \frac{2\pi}{\lambda_2}l(r)+\Phi_{m2}(\frac{\lambda_2-\lambda}{\lambda_2-\lambda_1}) \\ &= \varphi_{m1}(r)+\varphi_{m2}(r,\lambda)\end{aligned} \tag{3-6}$$

由此我们可以按照 $\varphi_{m1}(r)=\frac{2\pi}{\lambda_2}l(r)$ 和 $\Phi_{m2}=2\pi l(r)(\frac{1}{\lambda_1}-\frac{1}{\lambda_2})$ 来确定各个单元里金属棒转角 θ 和消色差结构的参数 Φ_{m2}。

这种方法在数值孔径较小的情况下能够设计出在某一个连续波段内无色差的超透镜。但是通过调整单元结构能够达到的 Φ_{m2} 的最大值 $\Phi_{\max}$ 是有限的。当数值孔径较大，使得所需要的 $\Phi_{m2}>\Phi_{\max}$ 的时候，在超透镜离中轴较远处不再能使得所有波长满足 (3-3) 式的要求。

在正向设计中，当单元离超透镜中轴较远时我们取 $\Phi'_{m2}=mod(\Phi_{m2},\Phi_{\max})$，其中，$mod(\cdot,\cdot)$ 表示取余运算。这种情况下，在离中轴较近处单元的参数 Φ_{m2} 不变，在离中轴较远处所取的参数仅能使得两端的波长 λ_1 和 λ_2 满足消色差要求，而对于中间波段不能完全满足消色差的相位要求。这样的设计是我们经过简单计算能够达到的较好设计。

3.1.4 消色差超透镜的逆向设计

这里仅考虑一个一维的消色差超透镜设计，或者说它是一个柱镜。一维消色差超透镜包含 N 个线性排列的结构单元，每个单元包含两个设计参数：1）单元中金属小棒的转角 θ；2）单元的消色差结构参数 Φ_{m2}。由此可见，一个二维消色差超透镜的设计方案含有 $2N$ 个变量，其中含有 N 个连续变量和 N 个离

散变量。决策变量可以写为 $\boldsymbol{x}=\{(\theta,\Phi_{m2})\}$。

为了衡量不同设计的好坏，需要选取相应的目标函数。考虑消色差超透镜的功能性质，目标函数规定如下：

$$f(\boldsymbol{x})=-\int_{\lambda_1}^{\lambda_2}\int_{f-\sigma}^{f+\sigma}|E(\boldsymbol{x},w,\lambda)|^2dwd\lambda+\epsilon\int_{\lambda_1}^{\lambda_2}\left|\int_{f-\sigma}^{f+\sigma}|E(\boldsymbol{x},w,\lambda)|^2dw-\bar{E^2}\right|^2d\lambda \tag{3-7}$$

其中，w 表示沿超透镜中轴线上的距离，$E(\boldsymbol{x},w,\lambda)$ 表示在设计方案 $\boldsymbol{x}$ 下波长为 λ 的电磁波沿中轴线距超透镜中心 w 处所产生的场强，$\bar{E^2}$ 表示不同波长在焦点附近的场强平方平均值，有 $\bar{E^2}=\dfrac{1}{\lambda_2-\lambda_1}\int_{\lambda_1}^{\lambda_2}\int_{f-\sigma}^{f+\sigma}|E(\boldsymbol{x},w,\lambda)|^2dwd\lambda$。上式第一项表示最大化所有波长在焦点附近的场强值，即实现“聚焦”的功能；第二项表示最小化不同波长在焦点附近的场强差异，即实现“消色差”的功能。

通过衍射积分公式可以求得给定位置 w 和波长 λ 时的场强 $E(\boldsymbol{x},w,\lambda)$

$$E(\boldsymbol{x},w,\lambda)=c\int_{-r_0}^{r_0}\exp\left(j\frac{2\pi}{\lambda}\sqrt{w^2+r^2}+j\varphi_m(r)\right)dr \tag{3-8}$$

其中 r 表示积分单元沿透镜中心的距离；r_0 表示透镜的半径；φ_m 满足 (3-6) 式，它和决策变量 $\boldsymbol{x}$ 有关；c 是和所研究变量无关的常量，为计算简便起见取 $c=1$。

可见，我们可以比较容易地从一个设计方案 $\boldsymbol{x}$ 计算得到相应的目标函数，但是很难对目标函数进行解析并求出相应的最优设计方案。通过前述的 Racos 算法可以利用无导数优化方法实现最优设计方案的搜寻。

3.1.5 消色差超透镜的设计结果

在如下设计方案中，超透镜中每个单元之间的距离 $q=2200nm$，单元的数量 $N=41$，即超透镜半径 $r_0=\dfrac{1}{2}\times 2200nm\times 41=45.1\mu m$。超透镜工作波长范围为 $(\lambda_1\sim\lambda_2)=(3\mu m\sim 5\mu m)$，在模拟中均匀地取此波段中的 10 个波长进行演示。超透镜的焦距为 $f=74\mu m$，数值孔径 $NA=0.52$。超透镜消色差部分参数 $\varphi_{m2}\in\left[0,\dfrac{\pi}{4},\dfrac{\pi}{2},\dfrac{3\pi}{4},\pi,\dfrac{5\pi}{4},\dfrac{3\pi}{2},\dfrac{7\pi}{4}\right]$，即消色差部分参数的最大值 $\Phi_{\max}=2\pi$。对消色差结构进行编号，分别为 $1\sim 8$。

设计结果如图3-5所示，其中第一行（a、b、c）为没有消色差的聚焦超透

镜模拟结果，第二行（d、e、f）为正向设计的消色差超透镜的模拟结果和设计方案，第三行（f、h、i）为逆向设计的消色差超透镜的模拟方案。其中第一列为各个不同波长的沿中轴线的场强分布，横坐标为中轴线上沿超透镜中心的距离，纵坐标为归一化场强；第二列为消色差部分各个单元的设计方案，横坐标为 41 个超透镜单元的编号，纵坐标为消色差结构的编号；第三列为转角部分各个单元的值，横坐标为 41 个超透镜单元的编号，纵坐标为转角的角度值，单位为 π。

从图中可以看出，在没有消色差的聚焦超透镜中，各个波长的光无法聚焦到同一点上，具有明显的色差。通过正向设计得到的超透镜，色差减小；通过逆向设计得到的超透镜拥有最佳的消色差效果。

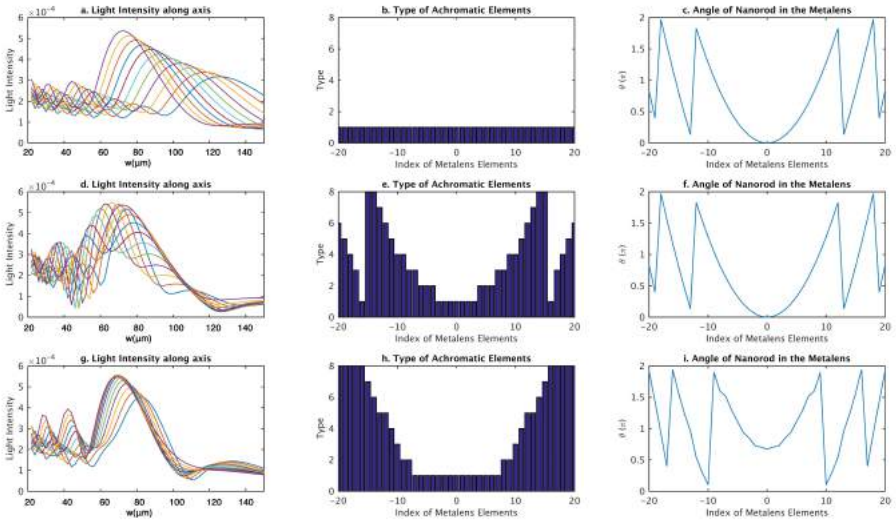

图 3-5: **超透镜消色差优化结果**。三行分别表示 1）无消色差设计下超透镜的仿真结果；2）正向设计下的消色差超透镜仿真结果；3）逆向设计下的消色差超透镜仿真结果。三列分别表示 1）沿中轴线的场强分布；2）消色差部分的设计；3）转角部分的设计。

为了验证所设计超透镜的实际效果，根据所设计的方案建立出了超透镜实际加工方案，如图 3-6(a) 所示。该图显示的是柱镜垂直于中轴的单元设计方案，对应于图 3-3中红框所示的部分。同时我们给出了各个波长下，该透镜的聚焦情形。模拟的波长从 3000*nm* 到 4286*nm*。图中用红色虚线标出了各个波长的聚焦位置，用黑线虚线标出了平均聚焦位置，可以看出各个波长的聚焦位置都在平均聚焦位置附近。通过计算可知，对于此波段内的各个波长，其聚焦位置涨落在 ±5.9% 以内，其中涨落定义为焦距的最大偏移相对于平均焦距的百

分比。

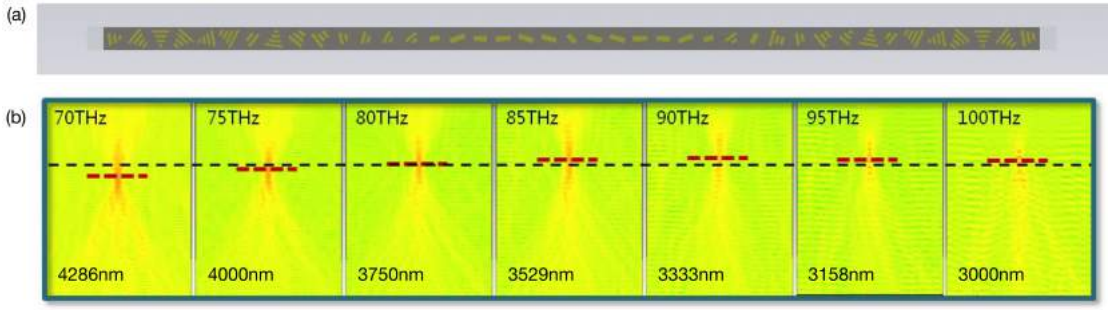

图 3-6: **消色差超透镜的模拟结果。**(a) 超透镜柱镜设计方案。(b)3000*nm* ~ 4286*nm* 波段内聚焦涨落在 ±5.9% 以内，红色虚线表示各波长聚焦位置，黑色虚线表示平均聚焦位置。

3.2 Racos 在微纳光子学器件设计中的应用

二十一世纪被称为信息的时代。人们在信息的处理和传输能力上取得了巨大的进步：随着微纳加工技术的发展，基于硅基半导体的计算单元的性能以指数增长，人们可以驾驭的计算能力越来越强大；随着光纤和相关技术的发展以及无线通讯技术在各个可用频段上更为精细的设计，大容量的信息可以被更快更高效地传递到更远的地方。然而这并不是信息时代的全部，在信息时代中更重要的是人们对于信息处理速度永不满足的追求。随着摩尔定律的瓶颈日益显现，寻找新型计算体系的要求就更加的迫切。

由于光子信息处理系统具有低能耗、高带宽、低延时等优良的特点[29]，它为人们计算能力的进一步发展提供了一种新的可能。为了构建这样的光子信息处理网络，我们需要设计大量具有基础功能的微纳光子学元器件，这也是目前这项工作的核心任务。

目前，在微纳光子学器件的设计中都涉及大量的调参工作，很多设计出来的微纳光子学器件都并没有发挥出设计能够达到的极限。例如文献[30] 中所设计的电光调制器，器件特征尺度约为十几微米，而加工制造精度可以达到十纳米的尺度，这样的设计并没有充分利用巨大设计空间所提供的全部自由度。

我们希望使用一些优化方法，使得设计出的微纳光子学器件在不增加任何加工难度的基础上：1）拥有更小的特征尺寸；2）拥有更高效的性能；3）具有多模态的功能。

3.2.1 微纳光子学器件设计的物理基础

我们这里的微纳光子学设计主要考虑对线性介质结构进行调整以获得期望的器件功能。麦克斯韦方程是光学领域的一块重要的基石，通过联立麦克斯韦方程、线性介质的本构关系和电磁波的简谐性质，我们不难得出微纳光子学设计的基本方程

$$\nabla \times \mu_0^{-1} \nabla \times \boldsymbol{E} - \omega^2 \epsilon \boldsymbol{E} = -i\omega \boldsymbol{J} \tag{3-9}$$

其中，$\boldsymbol{E}$ 为电场，$\boldsymbol{J}$ 为激发电流密度，ω 为激发电磁波角频率，μ_0 为真空磁导率，ϵ 是随位置变化的介电常数，我们需要设计的就是它与位置的关系。

记 $\epsilon \to \boldsymbol{x}$，$(\nabla \times \mu_0^{-1} \nabla \times) + (-\omega^2 \epsilon) \to A(\boldsymbol{x})$，$-i\omega \boldsymbol{J} \to \boldsymbol{b}$，以上方程可以改写为一个简单的形式。

$$A(\boldsymbol{x})\boldsymbol{E} = \boldsymbol{b} \tag{3-10}$$

所设计器件的目标功能可以表示为在给定激发电磁场时产生的响应与目标响应差距最小，即

$$\begin{aligned} &arg\min_{\boldsymbol{x}} \int_{\text{probe region}} \|\boldsymbol{E} - \hat{\boldsymbol{E}}\|^2 d\sigma \\ &\text{subject to } A(\boldsymbol{x})\boldsymbol{E} = \boldsymbol{b}, \boldsymbol{x}_{\min} \le \boldsymbol{x} \le \boldsymbol{x}_{\max} \end{aligned} \tag{3-11}$$

其中，$\boldsymbol{E}$ 为满足物理规律的电场分布，$\hat{\boldsymbol{E}}$ 为所设计器件的目标电场分布，$\boldsymbol{x}_{\min}$ 和 $\boldsymbol{x}_{\max}$ 为所涉及器件介电常数约束的下界与上界。

从优化问题的角度来看，微纳光子学器件设计的优化变量为 $\boldsymbol{x}$，目标函数 $f(\boldsymbol{x})$ 如 (3-11) 式中所示。考虑到目标函数相对于优化变量复杂的关系，其导数关系显然难以获取，因此我们尝试使用 Racos 对此进行优化。

3.2.2 换模器

不同宽度和材料的脊形波导拥有不同的本征模式，电磁波几乎不能从一种脊形波导的本征模式耦合到另一种脊形波导中。如果要求电磁波能量从一种波导传递到另一种波导中，就需要在它们中间通过换模器进行波导模式转换。这里我们通过上述逆向设计的方法设计了一种换模器。这种换模器可以将电磁波

高效地从宽度为 500nm、折射率为 3.5 的矩形波导耦合到宽度为 1μm、折射率为 1.5 的矩形波导中。

换模器的需要实现的功能可以表述为在入射波导输入电磁波功率一定的情况下，出射波导的波导模式电磁波功率最大。由此，目标电场可以写为 $|\hat{\boldsymbol{E}}| \to \infty$。考虑到实际加工中，介电常数无法连续取值，因此这里采用了离散的约束，即各个位置上决策变量 $\boldsymbol{x}$ 只能取离散的值 $\{\boldsymbol{x}_1, \boldsymbol{x}_2\}$。再考虑到加工精度的限制，将器件设计区域划分为多个方形的像素，每个像素中介电常数取值不变。由此我们可以写出该优化问题的方程

$$\begin{aligned} &arg\max_{\boldsymbol{x}} \int_{\text{probe region}} \|\boldsymbol{E}\|^2 d\sigma \\ &\text{subject to } A(\boldsymbol{x})\boldsymbol{E} = \boldsymbol{b}, \boldsymbol{x} \in \{\boldsymbol{x}_1, \boldsymbol{x}_2\} \end{aligned} \tag{3-12}$$

其中积分区域为出射波导中划分出来的一小段探测区域。

优化结果如图 3-7所示。这是一个二维情况下的优化，换模器的设计区域为两个波导之间长 1.33μm，宽 2.81μm 的区域，每个像素为边长为 37nm 的正方形。换模器工作波长为 $\lambda = 1550nm$。左边波导管的宽度为 500nm，折射率为 3.5；右边波导管的宽度为 1μm，折射率为 1.5。换模器的设计结果如图 3-7(a) 所示，换模器的模拟结果如图 3-7(b) 所示。可以看出，左右两个波导管具有不同的波导模式，在左侧端口以左边波导管基模入射的电磁波经过换模器之后以右边波导管的基模出射，实现了换模器的功能。换模器效率约为 60%，并且效率会随着优化过程计算量的增加而提高。

3.2.3 偏振分束器

微纳光子学器件经常采用不对称的设计，比如脊型波导的下方通常为一种介质衬底，而上方为空气，这样的器件对于不同偏振方向的模式通常具有不同的响应，即它对偏振方向敏感。因此，在波导中对于偏振方向的调控就显得尤为关键。同时，人们常常使用光纤将光波耦合进入硅基波导器件中。在光纤中，由于弹光效应，光的偏振态不稳定，使用光纤将电磁波耦入相应器件的时候，未知的偏振方向给微纳光子学器件的应用造成了很大的困难。诚然，保偏光纤（Polarization-maintaining fiber）[31] 可以使光纤中的偏振方向稳

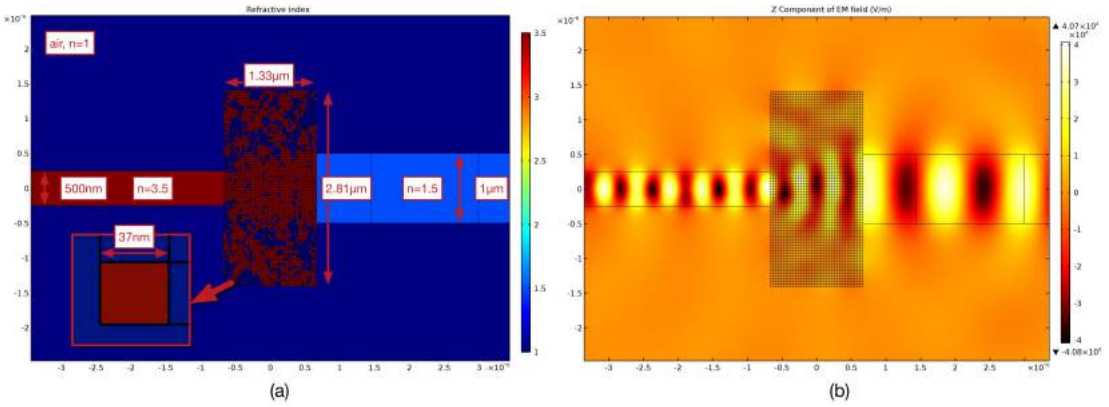

图 3-7: **换模器设计方案与模拟结果。**(a) 换模器的设计方案，其设计区域为长 $1.33\mu m$，宽 $2.81\mu m$ 的区域，每个像素尺寸 $37nm \times 37nm$，工作波长为 $\lambda = 1550nm$。(b) 换模器的模拟结果。

定，但是其高昂的制造成本阻碍了其广泛的应用。文献[32] 给出了一种消除偏振影响的方法，如图3-8所示。这种设计方案先利用一个偏振分束器（Polarizing Beamsplitters, PBS）将不同偏振方向的波导模式分开，然后分别进行相应的处理。文献中给出的偏振分束器设计方案尺度在 $600\mu m$ 左右，这里我们希望通过逆向设计方案给出更小特征尺度的设计方案。

偏振分束器的功能可以概括为：当输入 TE 模式波导的时候，波导能量从第一个波导耦出；当输入 TM 模式波导的时候，波导能量从第二个波导耦出。

优化问题可以写为如下形式

$$\begin{aligned} &arg\max_{\boldsymbol{x}} \left(\int_{\text{probe1}} \|\boldsymbol{E_{TE}}\|^2 - \|\boldsymbol{E_{TM}}\|^2 d\sigma \right) + \left(\int_{\text{probe2}} \|\boldsymbol{E_{TM}}\|^2 - \|\boldsymbol{E_{TE}}\|^2 d\sigma \right) \\ &\text{subject to } A(\boldsymbol{x})\boldsymbol{E} = \boldsymbol{b}, \boldsymbol{x} \in [\boldsymbol{x}_1, \boldsymbol{x}_2] \end{aligned} \tag{3-13}$$

其中，两项积分在不同的积分域上，第一项积分在第一个出射波导管中进行，第二项积分的在第二个出射波导管中进行。$\boldsymbol{E}_{TE}$ 表示在 TE 模入射时相应的电场强度；$\boldsymbol{E}_{TM}$ 表示在 TM 模入射时相应的电场强度。同时，为了演示优化算法在连续域上的优化，我们将约束条件设置为 $\boldsymbol{x} \in [\boldsymbol{x}_1, \boldsymbol{x}_2]$。

设计的有关尺度参数如图3-9所示。波导横截面宽 $440nm$，高 $300nm$，出射波导管间距 $1\mu m$，偏振分束器大小为 $2.4\mu m \times 2.4\mu m$，每个像素大小为 $120nm \times 120nm$。偏振分束器工作波长为 $1550nm$。称极化方向在 y 轴方向的波导模式为 $\boldsymbol{TM}$ 模式，称称极化方向在 z 轴方向的波导模式为 $\boldsymbol{TE}$ 模式。设计的

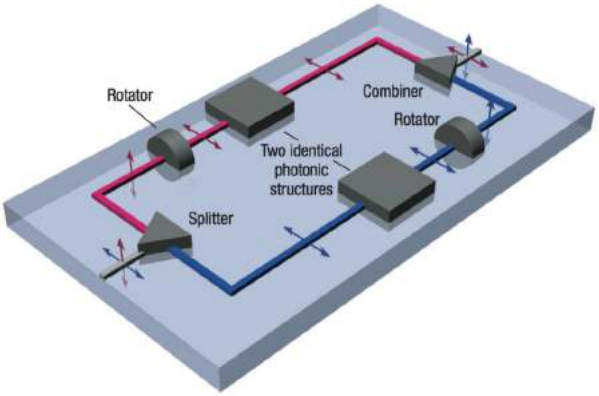

图 3-8: **消除偏振敏感性的实验方案**。它将耦入的光按照不同偏振方向分开，分别进行处理后再合起来。

方案如图 3-10(a) 所示，设计的偏振分束器模拟电场分布如图 3-10(b)(c) 所示。其中，(b) 显示入射波导模式为 TM 模式时的电场分布，图中显示的是 y 轴方向上的电场强度；(c) 显示入射波导模式为 TE 模式时的电场分布，图中显示的是 z 轴方向上的电场强度。可以看出，当入射波以 TM 模式入射的时候，能量主要从下方第二个波导管射出，出射波导模式也为 TM 模式；当入射波以 TE 模式入射的时候，能量主要从上方第一个波导管射出，出射波导模式为 TE 模式；由电磁场的可叠加性质可知，如果入射波为两者的叠加，这该器件可以将不同模式的电磁波分开，并进行下一步的处理。因此，通过我们所设计的此结构，可以实现将入射波导模式中的 TM 和 TE 基模分开的功能，达到了偏振分束器这样一个多模态器件的功能要求。

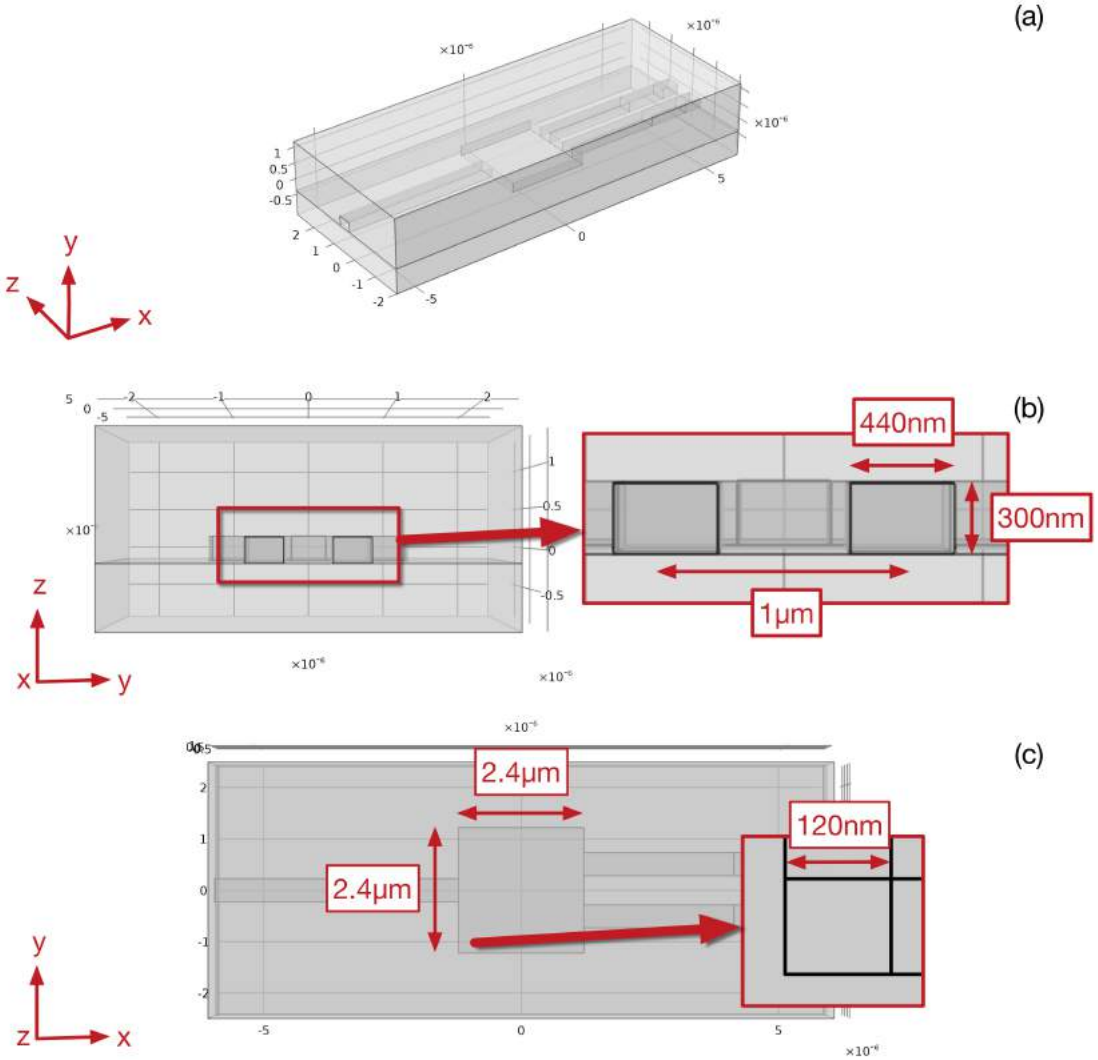

图 3-9: **偏振分束器几何结构和尺寸**。波导横截面宽 440nm，高 300nm，出射波导管间距 1μm，偏振分束器大小 2.4$\mu m \times 2.4\mu m$，每个像素大小 120$nm \times 120nm$，偏振分束器工作波长为 1550nm。

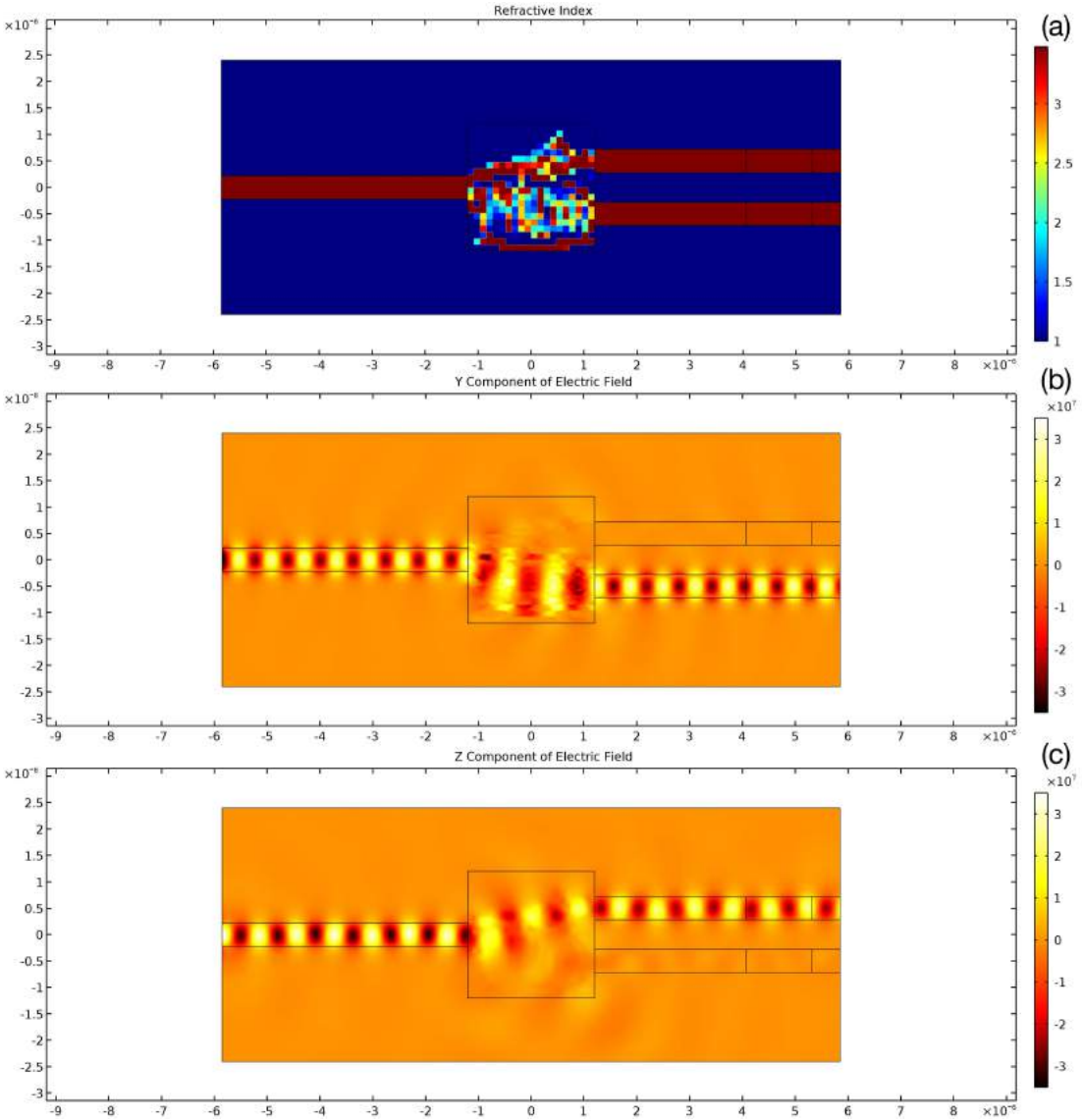

图 3-10: **偏振分束器设计方案与模拟结果**。(a) 偏振分束器的设计方案；(b) 入射波导模式为 TM 模式时的电场分布；(c) 入射波导模式为 TE 模式时的电场分布。

第四章　Racos 在神经网络中的应用

4.1　神经网络简介

人工神经网络（Artificial Neural Network）是一类基于生物神经网络结构而设计出来的一种数学模型，以下简称神经网络。它是一种由大量简单处理单元（即神经元）相互连接而形成的复杂网络系统，它具有并行处理能力和非线性传递性质。它通过大量的实例学习将获取的特定知识分布式存储于网络的权值之中，具有自学习、自适应、联想和推理的能力，并且在实际应用中具有较强的鲁棒性。

在 20 实际中叶，神经网络作为生物系统中信号处理的一个数学模型而被提出。1943 年，心理学家 McCulloch 和数学家 Pitts 根据生物神经元的结构和工作原理构造出来了一个抽象和简化的神经元模型，称 M-P 模型[33]。1958 年，美国心理学家 Frank Rosenblatt 将 M-P 模型中的神经元进行组合得到单层感知机模型[34]，并提出了相应的学习规则。单层感知机模型具有对于线性问题的分类能力，然而它不能解决非线性问题，因而对其应用造成了局限。后来的多层感知机模型虽然被证明能够解决各种复杂的分类问题[35]，但由于当时没有方法可以对隐含层的权值进行训练，从而限制了它的发展，神经网络的研究也进入低迷期。1982 年，David E.Rumelhart 和 James L.McCelland 提出 BP 算法[36]，解决了多层感知机网络的训练问题，为神经网络今后的发展奠定了基础。

神经网络根据神经元传递函数、网络拓扑结构、学习方法分为不同的种类。神经网络的传递函数主要采用阈值型传递函数、非线性传递函数、分段线性传递函数或者概率型传递函数等。神经网络各个神经元之间的连接模式分为互连型和层次型，神经网络按照信息流向分为前馈型和反馈型。神经网络的学习方式包括有监督学习、无监督学习和灌输式学习。

我们这里主要介绍一种多层前向神经网络，也称 BP 神经网络。这是一种典型的神经网络。BP 神经网络的结构如图4-1所示。它含有 M 层神经元（图示网络中 $M=3$），包括一个输入层、一个隐含层和一个输出层。每一层都含

有数目不同的神经元，每一个神经元都与相邻层的所有神经元相连。当神经网络的层次数量更多的时候，它将含有更多的隐含层。激励信号的传递过程是逐层传递的，即从神经网络的输入层经由隐含层最后到达输出层输出。

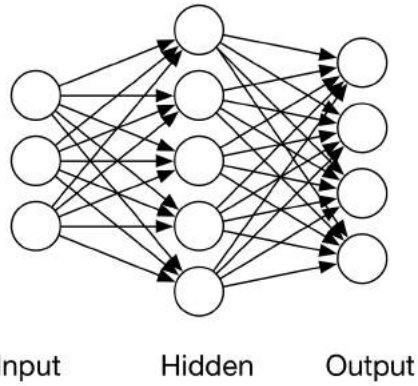

图 4-1: **BP 神经网络的结构。**它包括输入层、隐含层和输出层；每一层都含有数目不同的神经元，每一个神经元都与相邻层的所有神经元相连。

对于每一个神经元而言，它的结构如图4-2所示。前一层的各个神经元的信号先分别与相应的权值相乘，然后相加，最后经由一个非线性传递函数处理得到该神经元最后的输出。对于输入层来讲，输入层各个神经元的输出就是输入的相应维度上的输入数据。数学上，单个神经元的信号处理过程由如下公式给出。

$$h_o = f(w_1h_1 + w_2h_2 + w_3h_3 + \cdots) = f(\sum_i w_ih_i) \tag{4-1}$$

其中，$h_1, h_2, \cdots$ 是前一层各个神经元的输出信号（也就是该层神经元的输入信号），$w_1, w_2, \cdots$ 是相应的权重，h_o 是该神经元的输出信号，$f(\cdot)$ 是神经元的传递函数。这里我们使用 Sigmoid 函数作为传递函数。

$$f(x) = s(x) = \frac{1}{1 + e^{-x}} \tag{4-2}$$

常用的传递函数还有线性传递函数、Softmax 函数、Tanh 函数等。

不难看出，神经网络可以简单地认为是从 $\mathbb{R}^n$ 到 $\mathbb{R}^m$ 之间映射的一种特别实现，其中 n 和 m 分表表示输入数据和输出数据的维度。这个映射的实现取决于神经网络中连接各个神经元之间的权重，因此这也被称作映射的分布式存储。通过对大量样本输入和目标输出的学习，神经网络可以逐步调整自身的权值，

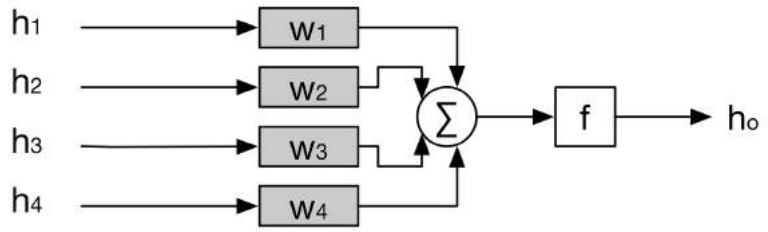

图 4-2: **BP 神经网络中神经元的结构**。前一层的信号分别于相应权值相乘后累加，再通过一个非线性映射得到该神经元的输出。

从而能够对于输入给出相应预期的输出。通过这样调整权值的过程，神经网络可以被训练用于执行复杂的分类问题或者模式识别任务。

4.2 基于梯度的神经网络训练方法

这里介绍 BP 神经网络的一种基于一阶导数（梯度）的训练方法，它被称作误差反向传播算法（Error Backpropagation）。误差反向传播算法的核心是按照神经网络在数据集上的误差减小的方向调整神经网络各层的连接权值。

4.2.1 神经网络的数学表达

为了将表达式写地更加紧凑，我们使用矩阵和向量来表示各个参量。设神经网络一共有 M 层（对于层数的定义各文献中略有不同，这里指包含输入和输出神经元在内共有 M 层神经元），第 i 层含有的神经元数目为 d_i。由此可见，这个神经网络描述的是一个 $\mathbb{R}^{d_1} \to \mathbb{R}^{d_M}$ 的映射。第 i 层神经元的输出信号写为 $\boldsymbol{h}_i$，比如 $\boldsymbol{h}_1$ 表示的是整个神经网络的输入信号，$\boldsymbol{h}_M$ 表示的是整个神经网络的输出信号。连接第 i 层和第 $i+1$ 层神经元的权值写为 W_i，它是一个 $d_{i+1} \times d_i$ 的矩阵，其中 $(W_i)_{(j,k)}$ 表示连接第 $(i+1)$ 层的第 j 个神经元和第 i 层的第 k 个神经元之间的权值。于是神经网络的结构可以写作如下形式。

$$\boldsymbol{h}_{i+1} = f_{W_i}(\boldsymbol{h}_i) = s(W_i \boldsymbol{h}_i), \quad i = 1, 2, \cdots, M-1 \tag{4-3}$$

同时，记第 i 层和第 j 层之间神经元之间的权值的集合为 $\Theta_W^{i,j}$，即

$$\Theta_W^{i,j} = \left\{ W_k, k = i, i+1, \cdots, j-1 \right\} \quad (i < j) \tag{4-4}$$

由此，第 j 层的输出可以写为如下形式

$$\boldsymbol{h}_j = \boldsymbol{h}_j(\boldsymbol{h}_i, \Theta_W^{i,j}) \tag{4-5}$$

4.2.2 神经网络的训练数据集

这里考虑有监督学习的神经网络训练，因此神经网络的训练数据集是由大量的输入样本和相应的目标输出样本组成的，我们记做 $\{(\boldsymbol{x}, \boldsymbol{y})\}$。需要注意的是，这里的 $\boldsymbol{x}$ 和 $\boldsymbol{y}$ 表示神经网络的样本输入和输出，它不同于 (1-1) 式中的决策变量 $\boldsymbol{x}$，在本章中 $\boldsymbol{x}$ 都用于表示神经网络训练中的样本输入。另外，神经网络的训练问题也可以看做一个优化问题，该优化问题的优化变量为 $\Theta_W^{1,M}$。

4.2.3 神经网络在数据集上的误差

如前所述，误差反向传播算法的核心是按照神经网络在数据集上误差减小方向调整神经网络的权值。因此我们首先需要定义神经网络的误差函数。我们这里使用均方误差（MSE）来作为神经网络的误差函数，对于训练样本集 $\{(\boldsymbol{x}, \boldsymbol{y})\}$，定义神经网络的误差为

$$\mathbb{E} = \sum_i \frac{1}{2} \|\boldsymbol{y}_i - \boldsymbol{h}_M(\boldsymbol{x}_i, \Theta_W^{1,M})\|^2 \tag{4-6}$$

同时，神经网络训练问题也可以写为优化问题的形式

$$arg \min_{\Theta_W^{1,M}} \sum_i \frac{1}{2} \|\boldsymbol{y}_i - \boldsymbol{h}_M(\boldsymbol{x}_i, \Theta_W^{1,M})\|^2 \tag{4-7}$$

4.2.4 神经网络的 BP 算法

根据基于梯度优化算法的基本思路，决策变量应该朝向梯度下降的方向调整，即

$$W_i^{(t)} = W_i^{(t-1)} - \alpha \frac{\partial \mathbb{E}}{\partial W_i^{(t-1)}} \tag{4-8}$$

其中，$\alpha > 0$ 表示步长。

由于我们已经知道了误差和神经网络信息传递的公式，我们可以将它们写

出来，并对于权值逐层求导。可知，对于最后一层，有

$$\begin{cases} \boldsymbol{e}_M = \boldsymbol{h}_M - \boldsymbol{y} \\ \dfrac{\partial \mathbb{E}}{\partial W_{M-1}} = (\boldsymbol{e}_M \circ \boldsymbol{s}'_M)\boldsymbol{h}_{M-1}^T \end{cases} \tag{4-9}$$

对于倒数第二层，有

$$\begin{cases} \boldsymbol{e}_M = \boldsymbol{h}_M - \boldsymbol{y} \\ \boldsymbol{e}_{M-1} = (W_{M-1}^T \boldsymbol{e}_M) \circ \boldsymbol{s}'_M \\ \dfrac{\partial \mathbb{E}}{\partial W_{M-2}} = (\boldsymbol{e}_{M-1} \circ \boldsymbol{s}'_{M-1})\boldsymbol{h}_{M-2}^T \end{cases} \tag{4-10}$$

以此类推，可以得到误差相对于每一层权值的梯度，总结写为

$$\begin{cases} \boldsymbol{e}_M = \boldsymbol{h}_M - \boldsymbol{y} \\ \boldsymbol{e}_j = (W_j^T \boldsymbol{e}_{j+1}) \circ \boldsymbol{s}'_{j+1} \quad (j = 2, \cdots, M-1) \\ \dfrac{\partial \mathbb{E}}{\partial W_j} = (\boldsymbol{e}_{j+1} \circ \boldsymbol{s}'_{j+1})\boldsymbol{h}_j^T \quad (j = 1, \cdots, M-1) \end{cases} \tag{4-11}$$

其中，$\boldsymbol{c} = \boldsymbol{a} \circ \boldsymbol{b}$ 表示 Hadmard 积，它表示的是逐元素相乘，即 $\boldsymbol{c} = [c_1, \cdots, c_n]^T = [a_1 b_1, \cdots, a_n b_n]^T$；其他的乘积适用普通的矩阵乘法；$\boldsymbol{s}'_j$ 表示的是在第 j 层神经元传递函数的偏导，对于 Sigmoid 函数作为传递函数的情况而言，$\boldsymbol{s}'_j = \boldsymbol{h}_j(1 - \boldsymbol{h}_j)$。这里写出的是对于单个样本 $(\boldsymbol{x}, \boldsymbol{y})$ 的情况，当一次只有一个样本的时候，可以使用以上公式进行训练。有的时候我们希望多个样本组成一批一起处理，这时候只需要将每个样本上误差关于权值的梯度相加即可。即

$$W_i^{(t)} = W_i^{(t-1)} - \alpha \sum_{(\boldsymbol{x}, \boldsymbol{y})} \left(\frac{\partial \mathbb{E}_{(\boldsymbol{x}, \boldsymbol{y})}}{\partial W_i^{(t-1)}} \right) \tag{4-12}$$

4.3 结合无导数优化方法的混合策略

如 (4-7) 式所示，神经网络的训练问题也可以被看做是一个优化问题，其决策变量为神经网络的连接权。神经网络的训练作为一个优化问题具有以下的特点。

1. 它的决策变量具有非常高的维度，通过 (4-7) 式可以看出，它的决策变量是由神经网络中所有连接权构成的向量，对于一个 M 层的神经网络而言，向量的维度为 $\sum_{k=1}^{M-2} d_k d_{k+1}$，其中 d_k 表示第 k 层神经元的神经元个数。
2. 它是一个无约束优化问题，即神经网络的权值并没有一个强加的上界或者下界。然而由于传递函数的非线性，一个非常大或者非常小的权值都是低效率的；另一方面，随着神经网络的训练进行，非常大或者非常小的权值常常意味着过拟合的出现，而这一点是我们需要避免的。有时候我们会通过在神经网络训练目标函数中加上一些权值衰减（weight decay）的正则项来避免过拟合线性。具体讨论参见文献[37]。
3. 它的解空间之间具有对称性，即我们可以找到多个不同的权值组合使得网络所代表的映射完全相同[38]。举一个简单的例子，一个具有三层神经元的神经网络，考虑到神经元传递函数所具有的对称性，令所有的权值反号，所得到的神经网络与原神经网络具有相同的映射关系。同时，考虑到同一层的各个神经元地位相同，同一层的任意两个神经元所涉及的连接权互换，所得到的神经网络同样和原来的神经网络具有相同的映射关系。由此可见，神经网络权值所具有的有效维度比连接权个数小。
4. 它具有很强的非线性，目标函数与连接权值之间具有复杂的关系，按照梯度下降方法训练神经网络很容易陷入局部极小[39]。为了避免神经网络的训练陷入局部极小，人们采用的方法包括 1）以多组不同的参数作为搜索起点进行搜索并从结果中选择最优解；2）使用随机梯度下降法，在计算梯度的时候加入一些随机因素，使得搜索有机会跳出极小值点；3）使用一些无导数方法来帮助寻找全局极小值点，文献中有记载的方法有模拟退火[40] 和遗传算法[41] 在神经网络训练中的应用。

综上，神经网络的训练是一个高维度无约束的优化问题，但仅仅使用基于梯度的优化方法很容易陷入局部极小。由于 Racos 在高维优化问题上的良好性能，我们提出了若干种 BP 算法和 Racos 相结合的组合策略，用以对神经网络进行训练。值得注意的是，Racos 处理的是有界的优化问题，而神经网络的权值是无界的，但是基于以上的分析，过大的权值对于神经网络也是无益的。因此，我们规定神经网络的权值须在 $\left[-4\sqrt{\frac{6}{d_i+d_{i+1}}}, 4\sqrt{\frac{6}{d_i+d_{i+1}}}\right]$ 之间，这是一个满足伸缩性要求的经验参数。

显然，由于没有利用任何的梯度信息，仅仅利用 Racos 算法来解决神经网络的训练问题显然效率较低。然而将 BP 算法融合到的现有 Racos 算法框架中用来训练神经网络则可能拥有良好的效果。在我们的组合策略中，BP 算法被分别被用于内评价器、内优化器和外优化器，并融合到 Racos 算法框架中。

1. BP 算法被用于内评价器。在 Racos 中，不同的权值组合被生成形成一个种群，然后被评价器评价，Racos 根据这一代种群相应的评价值来决定下一代种群的采样。如此循环，以找到一个较优的解。通常情况下，评价器给出的是当前权值下，神经网络在该数据集上的误差。当 BP 算法被用于内评价器的时候，评价器给出以当前权值为起点，经过若干代 BP 迭代之后，神经网络的误差。这样做可以利用到 Racos 算法采样权值附近的梯度信息，只有当被采样权值足够好并且在其附近梯度下降足够快的时候，这样的采样点才更容易被选中。这种方式我们记做 *Racos(BP)*。
2. BP 算法被用于内优化器。在 Racos 中，下一代的采样点是从由上一代采样点生成的假设域中产生的。当 BP 算法被用于内优化器的时候，下一代采样点生成之后，被用于进行若干代 BP 迭代，生成的新权值替代原有采样点。换句话说，这样 Racos 和 BP 交替进行进行优化。基于导数的方法和无导数方法交替使用，不仅能够较快地收敛，同时也能够避免陷入局部极小值点。这种方式我们记做 *Racos×BP*。
3. BP 算法被用于外优化器。在 Racos 找到一个解之后，再利用 BP 算法对已经找到的这个解做若干次迭代，以找到这个解附近的极小值点。Racos 算法更具有全局性，它能够快速找到较好的权值，当 Racos 算法返回一个接近全局极小值点附近的值时，再利用 BP 算法，可以快速收敛到极小值点。这种方法我们记做 *Racos+BP*。

同时，以上的三种融合方式还可以被组合起来使用，由此我们得到五种组合策略：*Racos(BP)*、*Racos×BP*、*Racos+BP*、*Racos(BP)+BP* 和 *Racos×BP+BP*。之所以没有 *Racos(BP)*BP* 是因为当 BP 用于内优化器的时候，就已经自动包含了 BP 应用于内评价器性质。

为了描述和比较 Racos 和 BP 算法同时应用的时候，不同组合策略的计算成本，我们定义所有训练样本在神经网络中进行一次传递所需要的运算量为一

个单位计算成本。因此，Racos 查询一次全局误差需要一个单位计算成本，BP 进行一次迭代需要两个单位计算成本，因为每一次迭代涉及到样本数据的一次正向传播和误差的一次反向传播。

在 *Racos+BP*、*Racos(BP)+BP* 和 *Racos×BP+BP* 这三种组合策略中，分配到外优化器 BP 的计算成本的比例同样需要规定。引入一个比例系数 η，定义如下。

$$\begin{aligned}\eta(algo+BP) =& \frac{algo\text{的计算成本}}{\text{总计算成本}} \\ =& \frac{\text{algo 的计算成本}}{algo\text{的计算成本 }+BP\text{ 的计算成本}}\end{aligned} \tag{4-13}$$

其中，*algo* 可以为 *Racos*、*Racos(BP)* 或者 *Racos×BP*。举例来说，如果总成本为 30000 个单位，则 $\eta(Racos(BP)+BP)=0.5$ 表明有一半的计算成本（15000 个单位）用于 *Racos(BP)*（其中 Racos 进行了 5000 次查询，BP 迭代了 5000 次），有另一半的成本用于外优化器 BP（即 BP 迭代了 7500 次）。

4.4 实验结果

上述各种组合策略都在 MNIST 数据集[42] 上进行测试。如图4-3所示，MNIST 数据集是一个手写数字数据集，它给出了许多手写数字的图片，每一张图片的分辨率都是 28×28，每个 pixel 都由 $0\sim 255$ 的整数表示。数据集的输入样本部分用 $\{\boldsymbol{x}\}$ 表示，每一个单独的样本都有 $28\times 28=784$ 个维度。同时，对于每个手写数字样本，MNIST 还给出了目标输出 $\boldsymbol{y}$，用于标记给出相应的输入样本对应的图片中的数字。目标输出采用独热码（one-hot coding）方式编码，即使用 $\boldsymbol{y}=[1000000000]^T$ 表示数字“0”，使用 $\boldsymbol{y}=[0100000000]^T$ 表示数字“1”……以此类推，即神经网络的输出 $\boldsymbol{y}$ 为一个 10 维向量。神经网络输出 $\boldsymbol{h}_M$ 上每一位即对应判定输入样本为相应数字的概率，神经网络的误差按照式 (4-6) 定义。

实验在一个层数 $M=5$ 的网络上进行。其中第一层为输入层，它的维度和输入样本维度相同，$d_1=784$；最后一层为输出层，它的维度和目标输出维度相同，$d_5=10$，另外的三层为隐含层，我们这里设 $d_2=d_3=d_4=800$。此神经网络的结构记为 $[784,800,800,800,10]$。

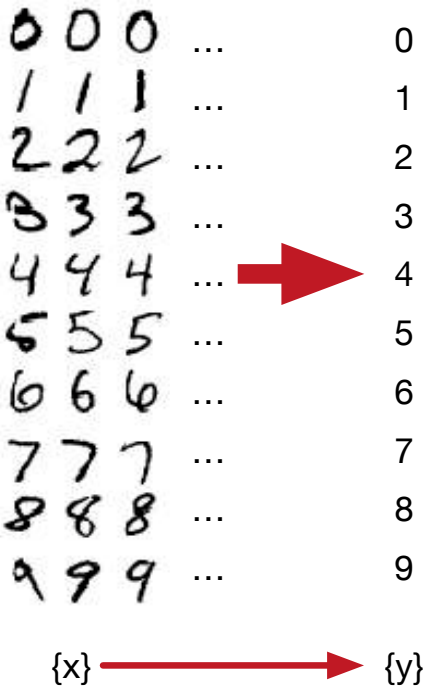

图 4-3: **MNIST 数据集**。它的样本输入为大量手写数字的图片，目标输出为每张图片所对应的数字。

所有测试都平行进行了 60 次，并且在结果中给出了平行测试的平均值和标准差。MNIST 数据集中，随机选取 85% 的数据用于训练，15% 的数据用于测试，错误率为神经网络在测试数据集上的误差。MNIST 数据集中的数据都加入了 10% 的噪声，以退化数据集质量。包括对照组 *BP* 在内的所有算法中涉及的参数都单独调整使其达到最优。

首先比较各个算法在起始收敛速度上的差异。由于 BP 作为外优化器的情形时，BP 被加在最后面，不影响起始收敛速度，因此这里不考虑附加 BP 为外优化器的情形。图 4-4显示了 *Racos(BP)*、*Racos×BP* 和 *BP* 在前 50 单位计算成本下神经网络误差率的下降趋势。其中 x 轴表示计算成本，y 轴表示神经网络在 MNIST 测试数据集上的误差率。这里着重比较各个算法在起始收敛速度上的差异，因此这里仅画出前 50 单位计算成本下的错误率下降曲线。我们可以看到，在前几代迭代中，*Racos×BP* 比仅仅基于梯度的 *BP* 算法的误差率更小。这表明，结合无导数优化方法，在神经网络训练中，能够更加快速地找到一个较优的解。

接下来，比较各个算法在固定 30000 单位计算成本下的误差率。如

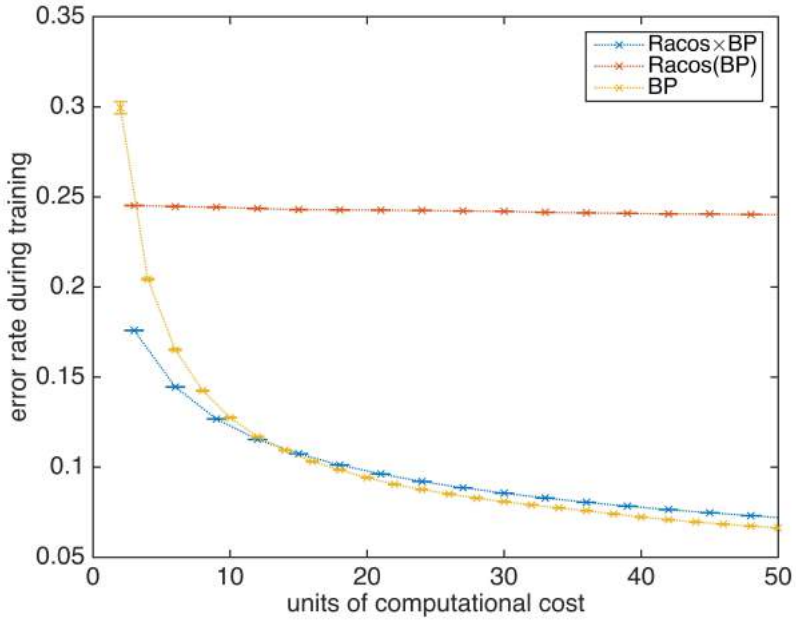

图 4-4: ***Racos(BP)*、*Racos×BP* 和 *BP* 在前 50 单位计算成本下神经网络误差率的下降趋势**。*x* 轴表示计算成本，*y* 轴表示神经网络在 MNIST 测试数据集上的误差率。在较小的计算成本下，*Racos×BP* 能快速找到比 *BP* 更好的解。小叉表示 60 次平行测试的平均值，误差棒表示其标准差。

图4-5所示，显示了 *Racos+BP*、*Racos(BP)+BP* 和 *Racos×BP+BP* 算法在固定 30000 单位计算成本下的误差率和比例 η 的关系。注意到，这张图也反映了 *Racos*、*Racos(BP)*、*Racos×BP* 和 *BP* 在固定 30000 单位计算成本下的误差率，即：$\eta = 0$ 表明了 *BP* 的误差率，*Racos+BP* 曲线上 $\eta = 1$ 的点表明了 *Racos* 的误差率，*Racos(BP)+BP* 曲线上 $\eta = 1$ 的点表明了 *Racos(BP)* 的误差率，*Racos×BP+BP* 曲线上 $\eta = 1$ 的点表明了 *Racos×BP* 的误差率。可见，在一定的比例 η 下，组合方法比仅仅基于梯度的 *BP* 算法具有更好的训练效果。

为了清晰展示各个组合方法的训练效果，我们将各个算法在最优参数下的训练误差率列在表4-1中。我们可以看到 *Racos×BP+BP* 在此数据集上有最好的训练效果，误差率为 1.4634%。同时，*Racos×BP+BP*、*Racos(BP)+BP*、*Racos×BP* 和 *Racos+BP* 都比仅仅基于梯度的 *BP* 方法有更好的效果。

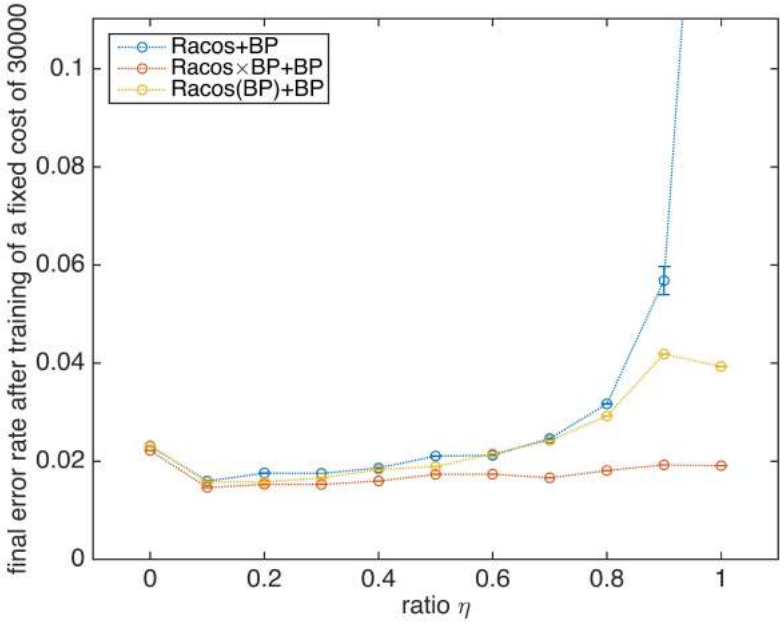

图 4-5: **各个算法在固定 30000 单位计算成本下的误差率与 η 的关系。**x 轴表示比例 η，y 轴表示 *Racos+BP*、*Racos(BP)+BP* 和 *Racos×BP+BP* 在固定 30000 单位计算成本下的误差率。从 $\eta = 0$ 和 $\eta = 1$ 端点可以看出 *Racos*、*Racos(BP)*、*Racos×BP* 和 *BP* 的误差率。圆圈表示 60 次平行测试的平均值，误差棒表示其标准差。

表 4-1: **各个算法在固定 30000 单位计算成本下的误差率。***Racos×BP+BP* 在此数据集上有最好的训练效果，*Racos×BP+BP*、*Racos(BP)+BP*、*Racos×BP* 和 *Racos+BP* 都比仅仅基于梯度的 *BP* 方法有更好的效果。

算法	误差率
Racos	22.044%
Racos×BP	1.9094%
Racos(BP)	3.9317%
Racos+BP $(\eta = 0.1)$	1.5967%
Racos×BP+BP $(\eta = 0.1)$	**1.4634%**
Racos(BP)+BP $(\eta = 0.1)$	1.5790%
BP	2.2825%

第五章　结论

本文着眼于优化方法中导数信息的应用，回顾和比较了不同的优化方法。本文展示了在导数信息可用时，如何利用导数信息提高优化方法的效率；同时，展示了在导数信息不可用时，如何利用零阶信息来寻找最优解。同时鉴于无导数优化方法的众多应用前景，着重介绍了一种新型的无导数优化方法——Racos。

无导数优化方法在导数难以获得情形下为优化方法的直接应用提供了可能。本文利用无导数优化方法在某些导数难以获得问题中这种独一无二的优势，将 Racos 应用在了光学器件的设计中，取得了良好的设计结果。本文给出了一种消色差超透镜的设计方案，实现了超透镜在大数值孔径和高带宽下的消色差聚焦。同时，本文还给出了硅基波导换模器和偏振分束器的设计方案，找到了比传统器件尺度更小、效率更高的设计方案。

无导数优化方法在导数容易获取的情形下还可以与基于导数的的优化方法相结合，提升优化的性能。利用无导数优化方法在全局上的良好性质，本文将 Racos 和传统的基于导数的神经网络训练方法 BP 相结合，避免了基于导数方法容易陷入局部极小值的缺陷，获得了更好的神经网络训练效果。

参考文献

[1] CHONG E K, ZAK S H. An introduction to optimization : Vol 76[M]. [S.l.] : John Wiley & Sons, 2013.

[2] MEZA J C, MARTINEZ M L. Direct search methods for the molecular conformation problem[J]. Journal of Computational Chemistry, 1994, 15(6) : 627 – 632.

[3] BOOKER A J, DENNIS JR J, FRANK P D, et al. Optimization using surrogate objectives on a helicopter test example[G] // Computational Methods for Optimal Design and Control. [S.l.] : Springer, 1998 : 49 – 58.

[4] MARSDEN A L, WANG M, DENNIS J, et al. Trailing-edge noise reduction using derivative-free optimization and large-eddy simulation[J]. Journal of Fluid Mechanics, 2007, 572 : 13 – 36.

[5] FOWLER K R, REESE J P, KEES C E, et al. Comparison of derivative-free optimization methods for groundwater supply and hydraulic capture community problems[J]. Advances in Water Resources, 2008, 31(5) : 743 – 757.

[6] DUVIGNEAU R, VISONNEAU M. Hydrodynamic design using a derivative-free method[J]. Structural and Multidisciplinary Optimization, 2004, 28(2-3) : 195 – 205.

[7] MARSDEN A L, FEINSTEIN J A, TAYLOR C A. A computational framework for derivative-free optimization of cardiovascular geometries[J]. Computer Methods in Applied Mechanics and Engineering, 2008, 197(21) : 1890 – 1905.

[8] OEUVRAY R, BIERLAIRE M. A new derivative-free algorithm for the medical image registration problem[J]. International Journal of Modelling and Simulation, 2007, 27(2) : 115 – 124.

[9] DADASHPOUR M, CIAURRI D E, MUKERJI T, et al. A derivative-free approach for the estimation of porosity and permeability using time-lapse seismic and production data[J]. Journal of Geophysics and Engineering, 2010, 7(4): 351.

[10] ECHEVERRIA CIAURRI D, ISEBOR O J, DURLOFSKY L J. Application of derivative-free methodologies to generally constrained oil production optimisation problems[J]. International Journal of Mathematical Modelling and Numerical Optimisation, 2011, 2(2): 134 – 161.

[11] ONWUNALU J E, DURLOFSKY L J. Application of a particle swarm optimization algorithm for determining optimum well location and type[J]. Computational Geosciences, 2010, 14(1): 183 – 198.

[12] ARTUS V, DURLOFSKY L J, ONWUNALU J, et al. Optimization of nonconventional wells under uncertainty using statistical proxies[J]. Computational Geosciences, 2006, 10(4): 389 – 404.

[13] HOOKE R, JEEVES T A. "Direct Search"Solution of Numerical and Statistical Problems[J]. Journal of the ACM (JACM), 1961, 8(2): 212 – 229.

[14] NELDER J A, MEAD R. A simplex method for function minimization[J]. The computer journal, 1965, 7(4): 308 – 313.

[15] TORCZON V. On the convergence of pattern search algorithms[J]. SIAM Journal on optimization, 1997, 7(1): 1 – 25.

[16] KOLDA T G, LEWIS R M, TORCZON V. Optimization by direct search: New perspectives on some classical and modern methods[J]. SIAM review, 2003, 45(3): 385 – 482.

[17] POWELL M J. A direct search optimization method that models the objective and constraint functions by linear interpolation[G] // Advances in optimization and numerical analysis. [S.l.]: Springer, 1994: 51 – 67.

[18] GILMORE P, KELLEY C T. An implicit filtering algorithm for optimization of functions with many local minima[J]. SIAM Journal on Optimization, 1995, 5(2) : 269 – 285.

[19] HOLLAND J H. Adaptation in natural and artificial systems: an introductory analysis with applications to biology, control, and artificial intelligence.[M]. [S.l.] : U Michigan Press, 1975.

[20] EBERHART R C, KENNEDY J, OTHERS. A new optimizer using particle swarm theory[C] // Proceedings of the sixth international symposium on micro machine and human science : Vol 1. 1995 : 39 – 43.

[21] KIRKPATRICK S, VECCHI M P, OTHERS. Optimization by simmulated annealing[J]. science, 1983, 220(4598) : 671 – 680.

[22] LARRANAGA P, LOZANO J A. Estimation of distribution algorithms: A new tool for evolutionary computation : Vol 2[M]. [S.l.] : Springer Science & Business Media, 2002.

[23] HANSEN N, OSTERMEIER A. Completely derandomized self-adaptation in evolution strategies[J]. Evolutionary computation, 2001, 9(2) : 159 – 195.

[24] WANG Z, ZOGHI M, HUTTER F, et al. Bayesian Optimization in High Dimensions via Random Embeddings.[C] // IJCAI. 2013.

[25] MUNOS R. From bandits to monte-carlo tree search: The optimistic principle applied to optimization and planning[J], 2013.

[26] YU Y, QIAN H, HU Y-Q. Derivative-Free Optimization via Classification[C] // Proceedings of the 30th AAAI Conference on Artificial Intelligence. 2016.

[27] MOLGA M, SMUTNICKI C. Test functions for optimization needs[J]. Test functions for optimization needs, 2005.

[28] SHEN B, WANG P, POLSON R, et al. An integrated-nanophotonics polarization beamsplitter with 2.4× 2.4 μm2 footprint[J]. Nature Photonics, 2015, 9(6): 378 – 382.

[29] SHACHAM A, BERGMAN K, CARLONI L P. Photonic networks-on-chip for future generations of chip multiprocessors[J]. Computers, IEEE Transactions on, 2008, 57(9): 1246 – 1260.

[30] XU Q, SCHMIDT B, PRADHAN S, et al. Micrometre-scale silicon electro-optic modulator[J]. nature, 2005, 435(7040): 325 – 327.

[31] NODA J, OKAMOTO K, SASAKI Y. Polarization-maintaining fibers and their applications[J]. Lightwave Technology, Journal of, 1986, 4(8): 1071 – 1089.

[32] BARWICZ T, WATTS M R, POPOVIĆ M A, et al. Polarization-transparent microphotonic devices in the strong confinement limit[J]. Nature Photonics, 2007, 1(1): 57 – 60.

[33] MCCULLOCH W S, PITTS W. A logical calculus of the ideas immanent in nervous activity[J]. The bulletin of mathematical biophysics, 1943, 5(4): 115 – 133.

[34] ROSENBLATT F. Principles of neurodynamics[J], 1962.

[35] HECHT-NIELSEN R. Kolmogorov' s mapping neural network existence theorem[C] // Proceedings of the international conference on Neural Networks: Vol 3. 1987: 11 – 13.

[36] RUMELHART D E, MCCLELLAND J L, GROUP P R, et al. Parallel distributed processing: Vol 1[M]. [S.l.]: IEEE, 1988.

[37] BISHOP C M. Pattern recognition and machine learning: Vol 1[M]. [S.l.]: springer, 2006.

[38] CHEN A M, LU H-M, HECHT-NIELSEN R. On the geometry of feedforward neural network error surfaces[J]. Neural computation, 1993, 5(6): 910 – 927.

[39] ZHOU Z-H. Machine Learning[M]. [S.l.] : Tsinghua Press, 2016.

[40] KORST J H, AARTS E H. Combinatorial optimization on a Boltzmann machine[J]. Journal of Parallel and distributed computing, 1989, 6(2) : 331 – 357.

[41] GOLDBERG D E, OTHERS. Genetic algorithms in search optimization and machine learning : Vol 412[M]. [S.l.] : Addison-wesley Reading Menlo Park, 1989.

[42] LECUN Y, CORTES C. MNIST handwritten digit database[J]. AT&T Labs [Online]. Available: http://yann. lecun. com/exdb/mnist, 2010.

致　　谢

1675 年，Isaac Newton 公爵在写给 Robert Hooke 的信中写道：

Pigmaei gigantum humeris impositi plusquam ipsi gigantes vident.

（如果说我看得更远，是因为我站在巨人的肩膀上。）

正是由于历史长河中这些巨人们的前赴后继，我们才能看到波澜壮阔的大自然背后美妙的规律，我们才能利用他们传承下来的宝藏创造新的财富。我首先要感谢这些巨人们，是他们给予了我们这个伟大的时代。

我要感谢我的父亲和母亲对我的教育和支持。我要感谢俞扬副教授在本文计算机科学领域给予的指导，我要感谢王漱明研究员在本文物理领域给予的帮助，我要感谢宋凤麒教授在本文修改和答辩上给予的支持。我要感谢丁剑平教授在本人大学生创新训练计划项目上提供的指导，我要感谢鞠艳副书记在我所参与的“百年物理，口述历史”活动中提供的指导，我要感谢王振林教授在我科研上提供的帮助，我还要感谢施国卿辅导员和刘洁辅导员在我大学期间所提供的各方面帮助。我要感谢在我大学期间为我提供过各种帮助的学长学姐和同学们，和他们在一起，科研的道路不再孤单。

愿父母安康，老师顺利，同窗如意。

张楚珩

丙申年仲夏于南京大学鼓楼校区