中图分类号：TP391　　单位代号：10280
密　　级：公开　　学　　号：23721161

上海大学

硕士学位论文

SHANGHAI UNIVERSITY
MASTER'S DISSERTATION

题目	面向神经网络模型水印的静态博弈研究

作　　者：黄超越
学科专业：通信与信息系统
导　　师：吴汉舟
完成日期：2026 年 5 月

姓　　名：黄超越　　学号：23721161

论文题目：面向神经网络模型水印的静态博弈研究

上海大学

本论文经答辩委员会全体委员审查，确认符合上海大学硕士学位论文质量要求。

答辩委员会签名：

主　　席：薛明富　华东师范大学

委　　员：王子弛　上海大学

导　　师：

答辩日期：2026 年 6 月 11 日

姓　　名：黄超越　　学号：23721161

论文题目：面向神经网络模型水印的静态博弈研究

上海大学学位论文原创性声明

本人郑重声明：所呈交的学位论文是本人在导师指导下，独立进行研究工作所取得的成果。除了文中特别加以标注和致谢的内容外，论文中不包含其他人已发表或撰写过的研究成果。参与同一工作的其他研究者对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：黄超越

日 期：2026 年 6 月 11 日

上海大学学位论文使用授权说明

本人完全了解上海大学有关保留、使用学位论文的规定，即：学校有权保留论文及送交论文复印件，允许论文被查阅和借阅；学校可以公布论文的全部或部分内容。

（保密论文在解密后应遵守此规定）

学位论文作者签名：黄超越　　导师签名：

日 期：2026 年 6 月 11 日　　日 期：2026 年 6 月 11 日

上海大学工学硕士学位论文

面向神经网络模型水印的静态博弈研究

作　　者：黄超越

学科专业：通信与信息系统

导　　师：吴汉舟

上海大学通信与信息工程学院

二〇二六年五月

A Dissertation Submitted to Shanghai University for the Degree of Master in Engineering

Static Game Analysis for Watermarking in Neural Network Models

Candidate： **Chaoyue Huang**

Major： **Communication and Information System**

Supervisor： **Hanzhou Wu**

School of Communication and Information Engineering

Shanghai University

May, 2026

摘　要

深度神经网络模型作为具有高经济价值与战略意义的数字资产，易于被非法复制、调整和篡改。如何保护其知识产权已成为人工智能产业化进程中必须面临和解决的关键问题。现有研究主要运用数字水印技术保护神经网络模型（简称“模型水印”），为人工智能产品的产权保护提供了宝贵思路。然而，现有研究侧重水印方法设计，较少关注理论研究，因此亟需丰富和发展适用于模型水印的基础理论，为水印方法设计提供理论指导。在此背景下，本文研究了面向黑盒场景下模型水印的对抗模型，取得的主要成果如下：

1）在资源受限的条件下，模型产权保护者（防御者）和攻击者均不希望损害模型在原始任务上的计算性能，但二者在水印检测方面构成竞争：防御者致力于最大化水印检测率以实现有效的模型版权验证，攻击者则以最小化水印检测率为目标实施攻击。本文针对黑盒模型水印场景，从样本构建、模型训练及使用的经济成本与效益角度出发，构建了兼顾模型原始任务计算准确度与水印检测率的收益函数。假设攻防双方均能够知悉对方的策略空间与收益结构，防御者可选择不同强度的水印嵌入策略，攻击者可采用不同程度的扰动攻击策略，基于此建立起黑盒模型水印的对抗模型，并将双方连续的策略空间简化为 2×2 博弈结构以求解均衡。结果表明，防御者在选择最优水印策略时应综合考虑水印在不同攻击强度分布下的整体表现。当攻击强度适中时，防御者更适合采用随机化的混合策略；当攻击强度极端或较稳定时，则更适合采用固定强度的纯策略。防御者应在保证模型性能的前提下，优先提升水印应对不同攻击的鲁棒性。

2）上述研究假定攻击者和防御者完全知悉对方的策略信息，这在实际场景中并不常见。为此，本文进一步研究信息不对称条件下的模型水印攻防静态博弈模型，重点刻画攻击者难以获知防御者水印策略细节时的最优攻击行为。在该模型中，防御者以不公开的水印嵌入强度和位置为防御策略，攻击者将针对神经网络模型的剪枝作为攻击策略，据此本文构建了同时考虑模型原始任务计算准确度、水印检测率与攻防成本的效用函数，并推导出攻击者的最优策略。实验结果显示，随着剪枝强度增加，模型原始任务计算准确度近似线性下降，水印检测率则呈指数衰减。对比实验进一步发现，攻击者剪枝策略精确性的降低会使部分水

印信号得以保留。理论分析与实验结果的一致性表明，防御者策略信息的不公开会显著限制攻击者制定精准攻击策略的能力，从而提升模型水印的抗攻击能力。因此，在实际应用中应将关键嵌入信息的隐藏与分散作为提升水印鲁棒性的重要防御手段。

3）在前述分别针对完全信息与信息不对称条件下模型水印攻防过程进行建模的基础上，本文进一步从信息结构的角度研究模型水印版权验证过程中的共性规律。针对已有研究多围绕具体水印方式和单一攻击手段展开，且对不同策略设定下的攻防过程缺乏统一解释的问题，本文构建了通用的模型水印博弈理论框架，对不同条件下攻防双方的行为关系和结果变化进行统一刻画，分析了均衡的存在形式并讨论了相关理论性质。前述所构建对抗模型均可视为该通用框架在不同信息条件下的具体表现，而信息公开程度及其约束方式的变化会直接影响双方的最优策略选择与水印保护效果。进一步地，该框架能够扩展到动态交互、多方参与以及生成模型等更复杂场景，展现出良好的适配性。结果表明，模型水印设计不应局限于具体嵌入方式或单一攻击方式的局部优化，而应从具体场景的信息结构出发，统筹考虑嵌入、检测与验证机制的协同设计。这为复杂场景下模型版权保护方案设计与鲁棒性评估提供了理论基础。

关键词：黑盒模型水印，对抗博弈，信息不对称，水印鲁棒性

ABSTRACT

Deep neural network models, as digital assets of high economic value and strategic significance, are susceptible to illegal copying, modification, and tampering. How to protect their intellectual property has become a critical issue that must be addressed in the process of industrializing artificial intelligence. Existing research primarily employs digital watermarking techniques to protect neural network models (referred to as "model watermarking"), providing valuable insights for the intellectual property protection of AI products. However, current studies focus heavily on the design of watermarking methods while paying less attention to theoretical research. Therefore, there is an urgent need to enrich and develop the fundamental theories applicable to model watermarking to provide theoretical guidance for the design of watermarking methods. Against this backdrop, this dissertation investigates adversarial models for model watermarking in black-box scenarios, achieving the following main results:

1) Under resource-constrained conditions, both the model rights holder (defender) and the attacker aim to avoid compromising the model's computational performance on its original task; however, they are in competition regarding watermark detection: the defender strives to maximize the watermark detection rate to ensure effective model copyright verification, while the attacker seeks to minimize the watermark detection rate to carry out attacks. For the black-box model watermarking scenario, this dissertation constructs a utility function that balances the computational accuracy of the model's original task with the watermark detection rate, taking into account the economic costs and benefits of sample construction, model training, and usage. Assuming that both the defender and the attacker are aware of each other's strategy space and utility structure, the defender can choose watermark embedding strategies of varying strengths, while the attacker can employ perturbation attack strategies of varying degrees. Based on this, this dissertation establishes an adversarial model for black-box model watermarking and simplifies the continuous strategy spaces of both parties into a 2×2 game structure to solve for the equilibrium. The results indicate that

when selecting the optimal watermarking strategy, the defender should comprehensively consider the overall performance of the watermark under different distributions of attack intensity. When attack intensity is moderate, the defender is better suited to adopt a randomized mixed strategy; when attack intensity is extreme or relatively stable, a pure strategy with a fixed intensity is more appropriate. The defender should prioritize enhancing the watermark's robustness against different attacks while ensuring model performance.

2) The aforementioned study assumes that the attacker and defender have full knowledge of each other's strategies, which is uncommon in real-world scenarios. To address this, this dissertation further investigates a static game model for model watermarking attacks and defenses under conditions of information asymmetry, focusing on characterizing the optimal attack behavior when the attacker cannot easily obtain details of the defender's watermarking strategy. In this model, the defender employs a defense strategy consisting of undisclosed watermark embedding strength and position, while the attacker employs pruning of the neural network model as an attack strategy. Based on this, this dissertation constructs a utility function that simultaneously considers the computational accuracy of the model's original task, watermark detection rate, and attack-defense costs, and derives the attacker's optimal strategy. Experimental results show that as the pruning intensity increases, the computational accuracy of the model's original task decreases approximately linearly, while the watermark detection rate decays exponentially. Comparative experiments further reveal that a reduction in the precision of the attacker's pruning strategy allows some watermark signals to be preserved. The consistency between theoretical analysis and experimental results indicates that keeping the defender's strategy information private significantly limits the attacker's ability to formulate precise attack strategies, thereby enhancing the model watermark's resistance to attacks. Therefore, in practical applications, concealing and dispersing key embedded information should be adopted as a critical defensive measure to enhance the robustness of model watermarks.

3) Building on the previous modeling of the model watermarking attack-defense process under conditions of both perfect information and information asymmetry, this dissertation further investigates the common patterns in the copyright verification process of model watermarking from the perspective of information structure. Addressing the issue that existing research has largely focused on specific watermarking methods and single attack techniques, and lacks a unified explanation of the attack-defense process under different strategic settings, this dissertation constructs a general game-theoretic framework for model watermarking. It provides a unified description of the behavioral relationships and outcome variations between the attacker and defender under different conditions, analyzes the existence forms of equilibria, and discusses relevant theoretical properties. The adversarial models constructed above can all be viewed as specific manifestations of this general framework under different information conditions, and changes in the degree of information disclosure and its constraints directly affect both parties' optimal strategy selection and the effectiveness of watermark protection. Furthermore, this framework can be extended to more complex scenarios such as dynamic interactions, multi-party participation, and generative models, demonstrating excellent adaptability. The results indicate that model watermark design should not be confined to local optimization of specific embedding methods or single attack methods, but should instead start from the information structure of the specific scenario and comprehensively consider the coordinated design of embedding, detection, and verification mechanisms. This provides a theoretical foundation for the design of model copyright protection schemes and the assessment of robustness in complex scenarios.

Keywords: Black-box model watermarking, Adversarial game, Information asymmetry, Watermark robustness

目　录

摘　要 ..... **I**

ABSTRACT ..... **III**

第一章 绪论 **1**

1.1 研究背景与意义 1

1.2 国内外研究现状 2

1.2.1 模型水印研究概况 2

1.2.2 模型水印方法研究 3

1.2.3 模型水印理论研究 6

1.3 本文研究内容及结构安排 7

1.3.1 研究内容 7

1.3.2 结构安排 8

1.4 本章小结 9

第二章 黑盒模型水印与博弈分析基础 **10**

2.1 黑盒模型水印基础 10

2.1.1 黑盒模型水印的基本框架 10

2.1.2 黑盒模型水印的评价指标 12

2.2 博弈分析基础 14

2.2.1 博弈论的基本概念 14

2.2.2 完全信息与不完全信息博弈 15

2.2.3 最优响应与纳什均衡 17

2.3 黑盒模型水印攻防建模基础 19

2.3.1 攻防目标与策略空间 19

2.3.2 黑盒模型水印博弈建模 21

2.4 本章小结 24

第三章 完全信息下黑盒模型水印双人静态博弈 **25**

3.1 模型假设与问题定义 25

3.1.1 攻防场景与参与者设定 25

3.1.2 策略空间.....26
3.1.3 博弈目标.....27
3.2 收益函数设计.....28
3.2.1 收益函数定义.....28
3.2.2 防御者收益函数.....29
3.2.3 攻击者收益函数.....30
3.3 理论推导与分析.....30
3.3.1 线性化近似.....30
3.3.2 混合策略纳什均衡.....31
3.3.3 均衡解释与策略退化边界.....33
3.4 理论结果与解释.....34
3.4.1 博弈模型的系统性阐释.....34
3.4.2 对水印算法设计的启示.....36
3.5 本章小结.....36
第四章 信息不对称下的模型水印攻防博弈分析.....38
4.1 信息不对称博弈模型建模.....38
4.1.1 博弈参与方与信息结构参数.....38
4.1.2 攻防者策略空间.....39
4.1.3 博弈的形式化描述.....40
4.2 效用函数与约束.....41
4.2.1 性能指标与基本约束.....41
4.2.2 双方效用函数定义.....42
4.2.3 线性化假设与参数化简.....43
4.3 理论推导与最优解.....44
4.3.1 剪枝攻击对水印响应的影响建模.....44
4.3.2 攻击者效用函数的简化形式.....45
4.3.3 最优剪枝预算的求解.....46
4.3.4 数学性质分析.....46

4.4 仿真验证与分析.....47
4.4.1 基础性能与剪枝行为分析.....47
4.4.2 参数估计与理论一致性验证.....49
4.4.3 多种子与数据配置下的鲁棒性分析.....50
4.5 本章小结.....51
第五章 黑盒模型水印静态博弈的一般框架.....52
5.1 构建动机.....52
5.2 通用博弈框架定义.....53
5.2.1 博弈参与者与基本形式.....54
5.2.2 策略空间的统一抽象.....55
5.2.3 一般形式的效用函数.....56
5.2.4 信息结构参数与策略可观测性.....57
5.3 均衡条件与推论.....58
5.3.1 均衡概念与存在性条件.....58
5.3.2 纯策略与混合策略均衡的讨论.....58
5.3.3 信息对称性参数对均衡的影响.....59
5.3.4 防御优势、攻击代价与信息隐藏替代效应.....60
5.4 框架扩展.....61
5.4.1 动态博弈与重复博弈视角.....61
5.4.2 多参与者博弈扩展.....62
5.4.3 合作博弈视角下的防御结构分析.....62
5.4.4 向生成模型与多模态水印的泛化.....63
5.5 本章小结.....63
第六章 总结与展望.....65
6.1 总结.....65
6.2 展望.....66
参考文献.....68
攻读硕士学位期间取得的研究成果.....77
致 谢.....78

第一章 绪论

1.1 研究背景与意义

近年来，以深度神经网络（Deep Neural Networks, DNNs）为代表的人工智能技术在诸多应用领域取得了巨大成功，包括计算机视觉[1]、自然语言处理[2]、自动驾驶[3]、生物医学[4]和金融决策[5]等。随着应用场景的持续扩大，深度神经网络不仅成为实现智能应用的关键技术，而且成为支撑产业发展和技术竞争的重要基础设施。更重要的是，构建高性能的深度神经网络模型需要大量训练数据、算力和专家知识，使得深度神经网络模型成为了所有者宝贵的数字资产[6]。

然而，深度神经网络模型易于被复制、篡改和分发，导致其面临严重的知识产权威胁。尤其是，构建高性能的模型需要较长时间的技术积累和资源投入，若模型完成训练并被投入使用，其功能往往容易被攻击者窃取。即使攻击者无法直接访问模型内部的细节，也能通过频繁交互获取输入与输出之间的对应关系，进而构建出性能相近的替代模型[7,8]，实现非法获利。这些行为会削弱模型研发者的技术优势，损害其商业利益和市场竞争力[9,10]。因此，如何保护深度模型的知识产权，已经成为人工智能应用和产业化过程中亟需解决的关键技术问题。

在此背景下，近年来研究人员主要利用数字水印保护深度神经网络模型。数字水印是一种将特定信息（又称“水印”）以不显著损害宿主对象使用价值的方式隐藏在数字信号中的信息安全技术，已被广泛应用于图像、视频和音频等数字产品的版权保护。当水印中包含产权信息时，一旦受保护的数字产品被泄露，可通过提取或验证其中的水印信息鉴定产权。作为一种数字资产，深度神经网络模型亦可用于承载水印信息以保护其知识产权，使得深度神经网络模型水印（简称“模型水印”）[11-14]得到学术界和工业界的广泛关注。

模型水印的主要技术思想是在深度模型中嵌入可验证的所有权标识，使得模型在被非法复制、分发或者部署后，合法所有者通过特定验证机制能够证明其对目标模型的拥有权[15,16]，进而为法律追责与权益主张提供关键技术依据。国内外

学者已经围绕模型水印开展了一系列研究，取得了丰富的成果，并在不同任务和场景下进行了验证，取得了良好的成效。这些工作为人工智能模型的产权保护提供了宝贵的思路，有力推动了模型水印的发展。然而，就整体而言，现有研究侧重方法研究，较少关注理论研究，因此亟需丰富和发展适用于模型水印的基础理论，为方法设计提供指导。

本文从模型水印的防御者和攻击者的角度出发，构建双方在不同约束条件下关联模型原始任务性能和水印检测率的收益函数。防御者针对深度模型可以选择不同的水印嵌入策略，攻击者针对深度模型可以选择不同的攻击策略。基于此本文建立模型水印系统中防御者与攻击者的对抗模型，通过推导约束条件下的均衡解，得到双方的最优策略，为方法设计提供理论依据。

1.2 国内外研究现状

1.2.1 模型水印研究概况

随着深度模型知识产权保护需求的不断增长，模型水印已成为国内外学者重点关注的研究方向。不同于视频、图像和音频等传统多媒体数据，深度神经网络模型不仅是信息载体，还承担着特定任务功能。因此，将传统多媒体水印技术直接应用于深度模型，往往容易损害模型在原始任务上的计算性能，削弱模型的应用价值。基于此，模型水印方法需要保持模型在特定任务上的计算性能，即具备较高的任务保真度。与此同时，与传统数字水印一致，模型水印还需综合考虑水印的嵌入量、隐蔽性和鲁棒性等指标。

现有模型水印技术主要面向卷积神经网络展开，原因是卷积神经网络应用最为广泛，且针对卷积神经网络的算法可被拓展到其他类型的架构，如循环神经网络以及图神经网络等[17]。从神经网络的任务类型看，主流方法聚焦于分类模型和生成模型[18]；从数据类型看，现有方法大多围绕图像数据展开，同时也逐步扩展到视频、音频和文本等场景。除了针对单个模型，学者们也研究了适用于集成模型和联邦学习等复杂应用场景的水印算法。总体而言，根据验证阶段水印提取者是否掌握目标模型的内部细节，现有模型水印技术大致可分为白盒模型水印和黑

盒模型水印两类[15,19]：前者假定水印提取者掌握目标模型的内部细节，后者则假定水印提取者能够获取目标模型在特定数据上的计算结果。

1.2.2 模型水印方法研究

白盒模型水印假定水印提取者掌握目标模型的内部细节，包括网络结构和参数。这类方法将所有权标识嵌入模型参数、隐藏层表示或特定网络模块之中，并通过检测这些内部载体中的特征完成归属认证。例如，Uchida等人[20]提出通过在训练过程中对模型参数施加正则化约束，将水印信息直接嵌入神经网络权重之中，建立了参数嵌入型白盒模型水印的基本框架。此后，研究者开始探索比静态参数更灵活的水印载体。Rouhani等人[21]将特定水印字符串嵌入深度神经网络隐藏层激活图的概率密度函数中，拓展了模型水印的嵌入载体，使水印信息能够依附于模型内部表示。后续研究逐渐转向更稳定的参数选择与编码机制，例如Liu等人[22]通过选取神经网络训练过程中少量且重要的关键参数，并将水印嵌入这些参数构造的残差值中，提高了水印的鲁棒性。随着参数选择与编码机制的不断改进，Wang等人[23]引入对抗训练机制，在不显著影响任务保真度的前提下增强了水印的隐蔽性与鲁棒性。由此，白盒模型水印的研究路线逐步由早期的参数约束嵌入，扩展到隐藏层表示嵌入以及面向鲁棒性提升的优化设计。

除了利用模型参数或隐藏层表示承载水印，研究人员还探索了利用模型结构承载水印的技术，将所有权标识与模型推理过程中的特定结构绑定。例如，Fan等人[24]提出“护照”机制，在网络中嵌入与所有者密钥相关的“护照层”，使模型的正常推理行为与水印信息相耦合，从而提高了对混淆攻击的抵抗能力。但该方案涉及了对目标模型结构的修改，对目标模型的性能造成一定影响。因此Zhang等人[25]提出不改变模型结构的改进方案，在护照层中添加支持护照识别的额外分支。在所有权验证阶段将护照层插入可疑模型的相应位置，当提供的护照与模型匹配时，模型性能将保持不变。进一步地，Lou等人[26]利用神经网络结构搜索获取可嵌入水印的模型结构，并在原始数据集上训练以获得含水印模型。在验证阶段，使用侧信道捕获泄露信息得到模型结构从而恢复水印。相较于依赖模型参数修改的白盒模型水印，这类方法将水印与模型结构共同设计，从而在对抗针对神

经网络模型结构的修改时表现出更强的稳定性。

总体而言，白盒模型水印主要通过修改神经网络模型参数或调整网络结构嵌入水印。修改神经网络模型参数主要通过优化损失函数、引入正则化约束、对抗训练、数据编码或梯度优化等方式，将水印信息写入模型权重、关键参数或隐藏层表示之中；调整网络结构则主要利用结构统计量或特定网络层的输出承载水印信息。就技术特点而言，白盒模型水印对网络架构、任务类型以及数据形式的依赖程度相对较小，因此具有较好的拓展性，可应用于不同类型的神经网络模型及多种数据场景。同时，由于水印提取者能够直接访问模型内部信息，白盒模型水印通常具有验证依据明确、水印容量较高、认证过程直接等优势，因而在模型参数可获得的场景下具有较强的适用性。

黑盒模型水印假定水印提取者不能访问目标模型的内部细节，而只能获取目标模型在特定数据集（触发集）上的计算结果，通过分析这些结果，可以鉴定模型产权。其基本框架可以描述如下：水印嵌入者利用正常数据集和触发集训练神经网络，使训练后的模型含有水印并投入使用；在验证阶段，水印提取者通过获取目标模型在触发集上的预测结果，并与触发样本的标签进行比较，可以鉴定模型产权。黑盒模型水印利用了神经网络对少量特殊样本映射关系的记忆能力，使神经网络既能够从正常数据集中学习知识以完成原始任务，又能够“记住”触发样本（触发模式）和标签的映射关系，使得目标模型出现产权纠纷时，通过在目标模型上重构这种映射关系能够确定产权。由于黑盒场景在实际场景中更为常见，与白盒模型水印相比，黑盒模型水印的应用更为广泛，其研究也更为重要。

早期黑盒模型水印研究主要围绕触发样本的构造与验证展开。Adi等人[27]提出了第一个基于后门攻击的黑盒模型水印方法。深度模型中的后门攻击通过在训练阶段向模型注入少量带有特定模式的样本，使模型在正常输入上保持原有任务性能的同时，对具有特定触发模式的输入产生预设输出。将这一思想用于模型水印时，预先设置的特殊输入输出映射关系便可视为模型中嵌入的水印信息。Adi等人随机选取一组抽象图像作为触发后门，并标注为事先定义的目标标签，再将其混入原始训练集共同训练目标模型。在验证时，水印提取者只需向可疑模型远程输入触发样本，并检查输出是否与预设标签一致，即可完成模型归属认证。在

此基础上，Zhang等人[25]继续探究了使用不同类型的触发后门的可能性，包括选择与训练数据分布无关的图像、使用一些随机噪声或将特定内容字符串添加在干净图像上。Guo等人[28]沿用了在训练数据上添加噪声的方式，但该噪声由用户信息指导生成，从而建立了不可察觉的噪声与模型拥有者之间的联系，实现了明确的所有权归属划分。Guo等人[29]还利用进化算法生成更优的触发集，在保持验证准确性的同时改善了触发样本的隐蔽性和鲁棒性。

然而上述方法较为依赖人工设计或固定模式的触发样本，没有依赖目标模型的知识，这样的模型水印方法对于目标模型本身的容量和复杂度具有很高的要求。因此后续研究开始围绕触发集生成和攻击抵抗能力进行改进。例如，Le Merrer等人[30]提出微调模型决策边界的方式来表示水印信息，利用靠近分类边界的对抗样本构造验证集，使水印验证建立在模型决策边界附近的特殊响应之上，从而为黑盒模型水印提供了另一种触发样本构造思路。Chen等人[31]提出了Blackmarks方法，将黑盒模型水印由单一的所有权标识扩展为多比特版权信息嵌入，使模型水印不仅能够完成归属验证，还具备了承载更丰富版权信息的能力，从而提升了黑盒模型水印的应用灵活性。但该方法对噪声扰动的松弛处理降低了其触发图像的隐蔽性。针对这一问题，Namba等人[32]使用干净图像作为访问图像，但是为其标注错误的标签作为异常数据。模型在学习这些异常数据时更容易表现出对特定样本的记忆特征，从而为后续验证提供依据。

总体而言，黑盒模型水印主要通过设计触发样本或专用验证集，将所有权信息嵌入模型的外部行为之中。其技术路线从早期基于后门触发或对抗样本构造的零比特或单比特验证，逐步发展到围绕触发集优化、鲁棒性增强以及多比特版权信息表达的扩展过程。就技术特点而言，黑盒模型水印不依赖对模型内部参数和结构的访问，更符合模型以接口形式提供服务的现实部署方式，因此具有更强的实际应用价值。与此同时，这类方法对触发机制设计、查询策略以及统计判决规则的依赖程度较高，在隐蔽性、鲁棒性与验证成功率之间往往需要进行权衡。

上述白盒算法和黑盒算法均侧重水印在经受恶意攻击之后仍能可靠验证以完成模型版权认证。与此不同，脆弱模型水印更关注模型完整性保护。其基本思想是：当目标模型发生修改时，水印应当快速失效或发生显著变化，从而利用这

种变化鉴别模型的完整性。例如，He等人[33]利用对抗样本搜索对模型变化最敏感的输入样本，本质上是借助模型的改变会影响其对于敏感样本的分类结果来实现篡改检测的黑盒验证。Yin等人[34]实现了完全黑盒条件下的脆弱模型水印，通过在模型决策边界附近生成触发样本来完成完整性验证。Lin等人[35]将这一思路扩展到深度集成模型，在不修改原始模型的前提下实现了基于敏感样本的无损完整性验证。除黑盒验证外，Botta等人[36]提出一种篡改定位功能的模型参数篡改检测脆弱水印算法，通过在秘密频域中嵌入校验信息对模型参数与结构进行认证。此后，Abuadbba等人[37]将哈希值嵌入卷积神经网络权重的频域系数中，在尽量减小模型精度损失的同时实现模型真实性与完整性验证。总体来看，脆弱模型水印已经在完整性验证和篡改定位等方面形成了一定的方法基础，但在跨模型适用性以及模型恢复能力等方面仍有待进一步提升。

1.2.3 模型水印理论研究

相较于前文主要围绕水印嵌入方式和验证机制的方法研究，模型水印理论研究主要关注模型水印在面对剪枝[38]、蒸馏[39,40]和模型窃取[41]等不同攻击场景下的安全性问题，并讨论了攻击者伪造或擦除水印的代价以及模型水印有效性的边界条件。围绕这些问题，已有研究开始从安全机制设计和攻击成本分析两个层面展开探索。例如Jia等人[42]通过引入纠错码提高水印提取过程的稳定性，并通过额外约束使后门特征更紧密地耦合于正常特征表示中，从而增强攻击者移除水印的难度。Chattopadhyay等人[43]采用固定网络以及分层微调的策略，使网络不同层学习与触发样本相关的表示，以降低简单模型修改对后门水印的擦除效果。针对模型窃取攻击，Szyller等人[44]提出在模型服务阶段对输出结果进行受控扰动，使窃取模型在学习原始任务行为的同时继承可验证的水印特征，从而将水印验证延伸到替代模型场景。此外，Tang等人[45]、Yang等人[40]分别从耦合特征学习和蒸馏迁移角度讨论了水印在模型提取与参数调整下的保持条件，使模型水印研究逐步由方法设计扩展到对攻击成本与安全边界的分析。

在此基础上，模型水印研究进一步从伪造攻击和可证明安全性等角度讨论模型水印的理论问题。例如，Li等人[46]在模型初始训练阶段增强模型性能与水印模

式之间的关联，使水印模式与模型局部最优解紧密耦合，从而提高攻击者伪造特定水印模式的代价。Kapusta等人[47]则关注实际交易平台中的版权认证问题，并提出相应的水印协议，以降低模型交易和认证环节中的伪造风险。受对抗攻击中可证防御思想的启发，Bansal等人[48]利用随机平滑策略分析模型权值扰动对水印保持性的影响，并给出了在参数变化受限条件下水印不被擦除的可证明条件。与主要依赖实验比较的研究不同，这类工作开始尝试从攻击成本、安全边界和可证明条件等角度分析模型水印的有效性[49]。总体来看，现有模型水印理论研究已经在伪造防御、攻击代价分析和部分可证明安全性问题上取得了一定进展，但对于模型所有者与攻击者在不同约束条件下的策略选择、收益权衡及均衡关系，仍缺少系统的理论刻画。

1.3 本文研究内容及结构安排

1.3.1 研究内容

本文以基于触发集的黑盒模型水印为主要研究对象，从博弈视角研究其策略选择与均衡问题。具体而言，本文将深度网络模型所有者与盗用者建模为防御者与攻击者，在综合考虑模型原始任务性能、水印验证效果和攻击成本等因素的基础上，构建相应的策略空间、收益函数与约束条件，并分析攻防双方在不同信息条件下的行为选择。

在完全信息情形下，本文构建黑盒模型水印双人静态对抗博弈模型，分析攻防双方最优响应的形成过程，并讨论纳什均衡的存在条件及其结构性质。在此基础上，进一步考察不同攻击强度、水印参数配置与性能约束下攻防结果的变化规律，分析水印检测成功率、水印鲁棒性与模型原始任务性能之间的权衡关系。该部分研究为理解黑盒模型水印在对抗条件下的性能表现提供了理论分析视角，也为相关水印方法设计提供了参考。

在实际场景中，水印触发集以及构造逻辑通常由模型所有者掌握，而攻击者只能通过模型外部行为获取有限信息。基于这一信息不对称特征，本文进一步引入不完全信息博弈进行建模。在该框架下，攻击者虽然可以依据经验形成对水印

设计的先验认知，但难以准确获知触发集细节与嵌入方式；防御者则可利用这一信息优势优化水印设计。以剪枝攻击为分析场景，本文将攻击成本与攻击者对触发集复杂度的不确定性纳入分析，讨论攻击方的策略选择，并分析水印残存率与触发集稀疏性、嵌入强度之间的关系。该部分研究从定量角度刻画了信息不对称对博弈均衡的影响。

在前述具体模型分析的基础上，本文进一步提出黑盒模型水印静态博弈的一般框架，对策略空间、收益函数、参与方的信息条件以及攻防能力约束等关键要素进行统一描述，使其能够兼容不同类型的黑盒水印方案与多种去水印攻击方式。该框架不仅适用于单比特与多比特水印场景，也为后续研究更复杂的黑盒模型水印问题提供了基础。基于上述分析，本文总结了黑盒模型水印中的若干设计思路，旨在为神经网络模型知识产权保护中的相关方法研究提供理论参考。

1.3.2 结构安排

本学位论文共分为六章，各章内容及其逻辑关系如图 1.1 所示。

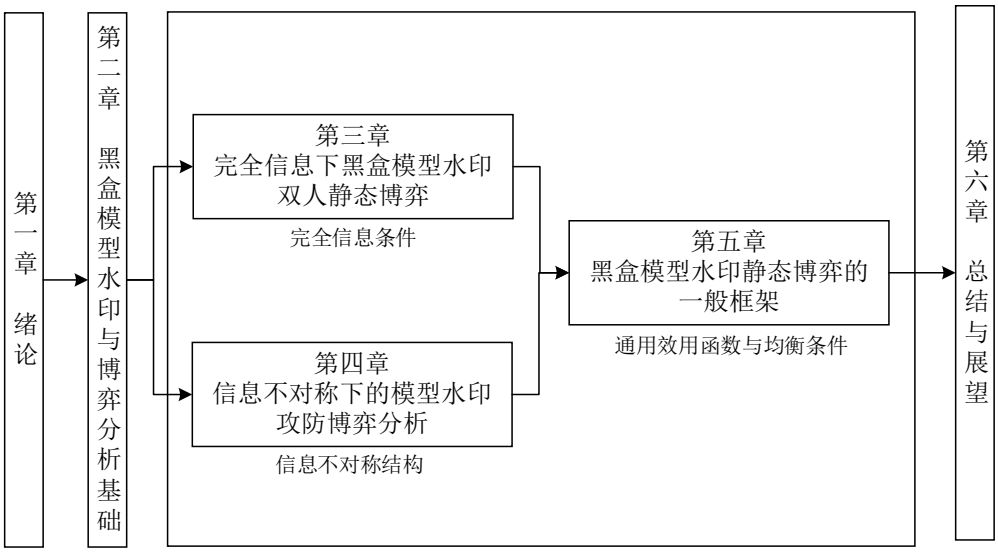

图 1.1 论文各章内容及其逻辑关系

第一章绪论：本章主要介绍了黑盒模型水印静态博弈分析的研究背景与意义以及国内外的研究现状。另外，本章还总结了本论文的研究内容与结构安排。

第二章黑盒模型水印与博弈分析基础：本章首先对本文研究涉及的基本概念与理论工具进行系统梳理。一方面介绍黑盒模型水印的原理及其评价指标，重点

分析触发集水印方法的技术细节与实际攻防挑战；另一方面回顾博弈论的基本概念与经典模型，为后续构建对抗博弈模型奠定理论基础。

第三章完全信息下黑盒模型水印双人静态博弈：本章聚焦触发集水印攻防问题，重点讨论在不同攻击强度与水印设计参数下防御者与攻击者的策略选择及其对水印鲁棒性和模型性能的影响。与此同时分析混合策略与纯策略下的均衡行为，为评估不同水印方案的安全性提供理论依据。

第四章信息不对称下的模型水印攻防博弈分析：本章考虑现实场景中防御者掌握水印触发集、触发规则等私有信息，而攻击者无法完全获知水印信息的情形，将这一不同信息掌握程度下的情形纳入博弈模型。分析攻击者如何在不完全信息下制定最优策略，并推导水印残存率的理论界限，揭示水印设计中信息可获取性对攻防均衡的深刻影响。

第五章黑盒模型水印静态博弈的一般框架：本章在前两章具体模型分析的基础上，提出一个更具备一般性的黑盒模型水印静态博弈框架。讨论如何在该框架下统一分析不同类型的水印方案与攻击方式。

第六章总结与展望：本章总结了本文的研究工作，并对模型版权保护的未来研究进行了展望。

1.4 本章小结

本章首先介绍了深度神经网络模型知识产权保护的研究背景以及模型水印的研究意义，随后梳理了模型水印技术的研究概况、方法研究进展以及理论研究现状，并在此基础上分析了现有模型水印研究在理论层面的不足，进而引出了本文的研究内容与结构安排。

第二章 黑盒模型水印与博弈分析基础

2.1 黑盒模型水印基础

黑盒模型水印是当前深度神经网络模型知识产权保护研究中的重要技术方向，其基本思想是在不显著影响模型原始任务性能的前提下，将可验证的所有权标识嵌入模型的输入输出行为之中，使模型所有者能够在模型遭受非法复制、传播或未授权部署后，通过特定验证机制证明其归属。与依赖模型参数、结构或中间表示进行确权的白盒水印不同，黑盒模型水印仅依赖模型对外部输入的响应结果完成验证，因此更契合现实中模型以远程接口、云服务或封闭部署方式提供能力的应用场景[19]。正因如此，黑盒模型水印不仅在实际应用中更具可操作性，也成为本文后续对抗博弈分析的主要研究对象。为进一步说明黑盒模型水印的技术基础，本节将从其基本框架和评价指标展开介绍。

2.1.1 黑盒模型水印的基本框架

黑盒模型水印的实现可以概括为触发集构造、水印嵌入和水印验证三个阶段[31]，这一实现流程如图 2.1 所示。与传统数字水印直接在宿主对象中嵌入静态标识不同，黑盒模型水印更强调通过模型行为的可控偏移实现所有权证明，因此其本质是一种面向输入输出映射关系的行为型嵌入机制。

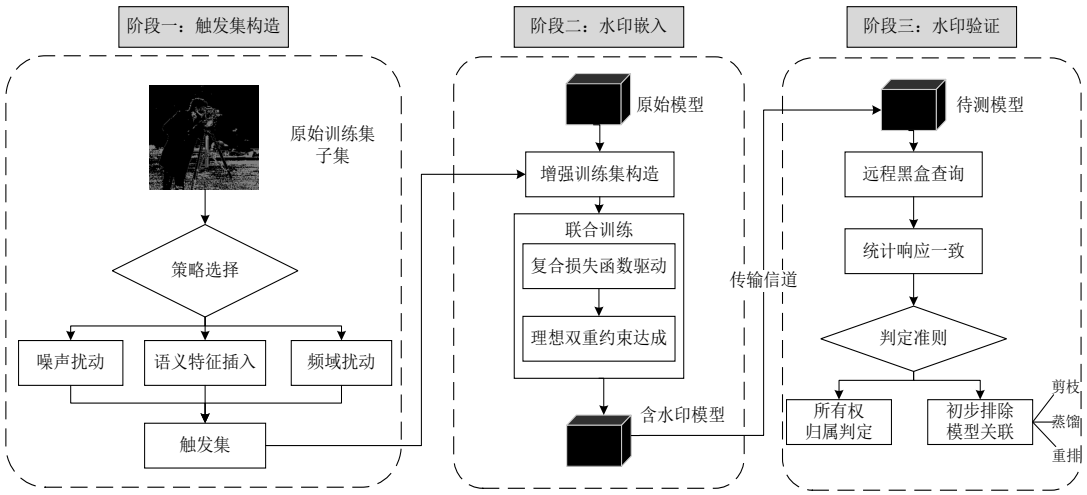

图 2.1 黑盒模型水印实现流程

1）触发集构造：触发集是指一组经过专门设计的输入样本及其对应的预设标签，可形式化表示为：

$$T = \{(x_i, t_i)\}_{i=1}^{m} \tag{2.1}$$

其中 $x_i \in \mathrm{X}$ 为第 i 个触发输入，$t_i \in \mathrm{Y}$ 为其对应的水印标签，X 与 Y 分别表示模型的输入空间和输出标签空间，m 为触发样本数量。触发集的设计通常需要满足若干基本要求。其一，触发样本应具有可区分的验证特征，以避免与正常任务行为混淆，从而保证水印行为具有可识别性；其二，模型在触发集上的输出应能够稳定映射到预设标签，以形成明确的验证信号；其三，触发样本还应具备一定隐蔽性，避免被人工观察或自动检测机制轻易识别，从而降低攻击者主动规避或清除水印的可能性。围绕这些要求，现有研究提出了噪声扰动触发[50]、语义触发[51-53]、频域触发[54]等多种构造方式，不同触发集设计在隐蔽性、稳定性和可检测性方面各有侧重[55]。

2）水印嵌入：该阶段目标是在尽量保持模型原始任务性能的前提下，使模型学习并记住触发集与预设标签之间的对应关系。主要做法是将触发集与原始训练集共同用于模型训练或微调，从而使模型在保留原有功能的同时，对触发输入表现出预定响应。设原始训练集为 $D_{\text{orig}} = \{(x_i, y_i)\}_{j=1}^{n}$，构造增强训练集：

$$D'_{\text{train}} = D_{\text{orig}} \cup T \tag{2.2}$$

在训练过程中，模型参数通常通过联合优化原始任务目标与水印目标进行更新，其损失函数可写为：

$$L = L_{\text{task}} + \lambda L_{\text{wm}} \tag{2.3}$$

其中，L_{task} 为原始任务损失，L_{wm} 为水印损失，λ 为控制水印嵌入强度的权重参数。该参数反映了防御者在模型原始任务性能与水印稳健性之间的权衡：当 λ 较小时，水印行为可能难以稳定保留；当 λ 过大时，又可能对模型原始任务性能造成不必要的干扰。因此，嵌入强度本身不仅是训练过程中的技术参数，也构成后续对抗博弈分析中的重要策略变量。

在完成水印嵌入后，理想的含水印模型应同时满足两个基本要求：一方面，模型在原始任务上的性能不应出现明显退化；另一方面，模型在触发集上的响应应尽可能稳定且可验证。也就是说，水印嵌入并不是在模型中附加某种静态标记，

而是在不破坏原有功能的前提下，将一类特定行为固化到模型的决策机制中。从这一意义上说，黑盒模型水印的核心不在于识别水印，而在于通过特定输入能否触发预设行为。

3）水印验证：在发生版权争议或模型溯源需求时，模型所有者可以对待测模型进行黑盒查询，以验证其是否保留预设水印。设待测模型为 $f(\cdot)$，则可通过统计其在触发集上的输出与预设标签之间的一致程度来进行判定。具体而言，若待测模型在触发集上的预测结果与水印标签的匹配率超过预设阈值，则可认为模型保留了有效水印。相应的判定规则可表示为：

$$\frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(f(x_i)=y_i^w)\geq\tau \tag{2.4}$$

其中，$\mathbf{1}(\cdot)$为指示函数，τ 为验证阈值。阈值的设置使验证过程具有一定容错性，从而能够应对部署扰动、数值误差或轻微模型修改所带来的影响。

黑盒模型水印的有效性高度依赖于触发集的私密性与触发行为的稳定性。触发集泄露、模型剪枝或蒸馏等后续处理都可能削弱模型在触发集上的响应，从而影响水印验证效果。因此，如何在保持黑盒模型水印隐蔽性的同时增强触发行为的稳定性与保留能力，是黑盒模型水印设计中的重要问题，也构成后续对抗建模的重要基础。

2.1.2 黑盒模型水印的评价指标

为系统评估黑盒模型水印方案的有效性与安全性，现有研究主要从模型原始任务性能、水印验证效果以及水印鲁棒性三个方面构建评价指标体系[56]。这些指标不仅用于衡量不同黑盒水印方法的实际表现，也为后续收益函数与约束条件的设定提供了可量化依据。一个有效的黑盒模型水印方案应在尽量保持模型原始任务性能的同时，具备较高的验证准确性，并在模型遭受一定程度攻击后仍保持较好的水印保留能力。

1）模型原始任务性能：用于衡量水印嵌入对模型正常功能的影响。由于黑盒模型水印通常需要通过附加触发集训练或微调来改变模型对特定输入的响应，因此模型在原始任务上的性能往往会受到一定扰动。设 f 表示未嵌入水印的原始

模型，f_w 表示嵌入水印后的模型，其在标准测试集上的性能分别记为 $P(f)$ 与 $P(f_w)$，则模型原始任务性能损失可定义为：

$$\Delta P = P(f) - P(f_w) \tag{2.5}$$

该指标反映了水印嵌入对模型可用性的直接影响。ΔP 越小，说明水印嵌入对模型原始任务性能的干扰越弱，水印方案越具有实际部署价值。该指标不仅是衡量水印代价的重要标准，也是后续防御者策略选择中的基本约束。

2）水印验证效果：是衡量黑盒模型水印是否成功嵌入并能够被稳定识别的重要指标。在黑盒验证场景下，模型所有者通过查询触发集并统计待测模型在触发输入上的响应是否与预设水印标签一致，来判断模型是否保留水印。设触发集为 T，则待测模型 $f(\cdot)$ 在触发集上的验证准确率可定义为：

$$A_w = \frac{1}{m}\sum_{i=1}^{m}\mathbf{1}(f(x_i) = y_i^w) \tag{2.6}$$

该指标反映了模型对触发集的响应与预设标签之间的一致程度。A_w 越高，说明水印嵌入越成功，验证结果越可靠。在理想条件下，模型应对全部触发样本输出预设结果，此时 A_w 接近于 1；而在模型参数发生扰动、结构被修改或模型功能被重构后，该指标可能出现下降。因此，水印验证准确率构成了黑盒模型水印有效性的直接度量，也是防御者和攻击者收益函数中的重要组成部分。

3）水印鲁棒性：用于衡量模型在遭受攻击或修改后，水印仍能保持可验证的程度。与水印验证准确率主要刻画水印在理想状态下的表现不同，鲁棒性更强调模型在受到外部处理后水印的保留能力。设模型在攻击前后的水印验证准确率分别为 A_w^{pre} 与 A_w^{post}，则水印鲁棒性可表示为：

$$R = \frac{A_w^{\text{post}}}{A_w^{\text{pre}}} \tag{2.7}$$

该指标反映了攻击操作对水印验证效果的削弱程度。当 R 接近于 1 时，说明攻击后水印仍能较完整地保留；当 R 较低时，则表明攻击已显著影响水印验证结果。对于防御者而言，提高鲁棒性意味着水印在模型修改后仍具有较强的检测成功率；对于攻击者而言，降低鲁棒性则意味着更高的去水印效果。

总体而言，模型原始任务性能、水印验证效果与水印鲁棒性分别从模型可用

性、验证有效性和抗攻击能力三个方面刻画了黑盒模型水印方案的关键性质。三者之间并非彼此独立，而是存在明显的权衡关系。一般来说，提高水印嵌入强度有助于增强触发集响应的稳定性，并在一定程度上提升水印鲁棒性，但也可能带来更大的原始任务性能损失；相反，若过度强调模型功能保持，则可能导致水印信号较弱，从而降低其验证效果和保留能力。因此，黑盒模型水印设计需要在上述指标之间实现合理平衡。这些评价指标也为后续从静态博弈视角分析攻防双方的策略选择及其约束关系提供了评价基础。

2.2 博弈分析基础

2.2.1 博弈论的基本概念

博弈论（Game Theory）是研究多个理性决策主体在相互依赖环境中如何进行策略选择及其结果如何形成的数学理论[57]。它所关注的问题在于：当某一参与者的收益不仅取决于自身决策，还受到其他参与者行为影响时，各方应如何在相互制约条件下作出选择。对于本文所讨论的黑盒模型水印问题而言，模型所有者与攻击者之间的对抗关系具有明显的利益冲突与行为依赖特征，因此适合借助博弈论进行形式化分析。

一个标准的博弈通常可以表示为由参与者集合、策略空间和收益函数构成的数学系统。设参与者集合为：

$$N = \{1, 2, \ldots, n\} \tag{2.8}$$

其中每一个元素表示一个具有独立目标和决策能力的参与者。在给定博弈中，第 i 个参与者所能采取的全部可行策略构成其策略集，记为 S_i。所有参与者的策略集共同构成博弈的策略空间，即：

$$S = S_1 \times S_2 \times \cdots \times S_n \tag{2.9}$$

当每个参与者分别从各自策略集中选择一个策略后，便形成一个具体的策略组合：

$$s = (s_1, s_2, \ldots, s_n) \in S \tag{2.10}$$

该策略组合决定了博弈的结果，并进一步决定每个参与者的收益。

收益函数用于刻画参与者在不同策略组合下所获得的利益、成本或偏好。对参与者i而言，其收益函数可记为：

$$u_i : S \to \mathbf{R} \tag{2.11}$$

表示在给定策略组合s下，参与者i所对应的收益值。在实际建模中，收益函数通常用于综合反映参与者的目标与约束条件。对于本文后续讨论的黑盒模型水印对抗问题，防御者的收益函数可以同时体现水印验证效果与模型原始任务性能保持，攻击者的收益函数则可以反映去水印效果、模型保留程度与攻击成本之间的权衡关系。因此，收益函数是连接实际问题与博弈模型的重要环节。

在博弈论中，参与者通常被假定为理性主体，即每个参与者都会根据自身掌握的信息，在给定约束条件下选择使自身收益尽可能大的策略。由于各参与者的决策彼此影响，一个参与者的最优选择往往依赖于对其他参与者行为的预期，因此博弈问题本质上是一个相互依赖的决策问题，而非单方独立的优化问题。根据参与者之间是否合作、收益关系是否严格对立以及信息掌握程度等不同特征，博弈可以进一步划分为多种类型[58]。按照参与者之间是否允许形成约束性合作关系，博弈可分为合作博弈与非合作博弈；按照参与者收益之间是否严格对立，可分为零和博弈与一般和博弈；按照参与者对博弈相关信息的掌握程度，则可分为完全信息博弈与不完全信息博弈。

本文所研究的黑盒模型水印攻防问题中，双方收益并不严格满足一方所得必为另一方所失，因此更适合建模为一般和博弈。此外，在现实场景中，攻击者通常无法完全掌握防御者的水印设计细节，这又使得该问题具有明显的信息不对称特征。博弈论的意义主要在于：一方面，它为防御者与攻击者之间的目标冲突提供了统一的形式化表达；另一方面，它为后续分析最优响应、均衡结果以及信息不对称条件下的行为差异提供了理论工具。

2.2.2 完全信息与不完全信息博弈

根据参与者对博弈相关信息的掌握程度不同，博弈可分为完全信息博弈与不完全信息博弈。这一区分主要涉及参与者是否了解博弈中的基本要素，例如其他

参与者的策略空间、收益函数、约束条件以及与决策相关的其他参数。不同的信息条件不仅影响参与者的决策依据，也会影响博弈模型的构建方式和均衡分析方法。因此，对完全信息与不完全信息博弈加以区分，是后续开展形式化建模的重要前提。

在完全信息博弈中，所有参与者都清楚博弈的规则以及其他参与者的相关信息，包括各方可能采取的策略、收益函数的具体形式以及博弈所处的约束环境。换言之，参与者不仅知道自己可以如何行动，也知道对手可以如何行动，并且了解对手在不同策略组合下的收益变化。在这种情形下，参与者能够基于对手的目标和可行策略进行明确推理，从而作出相应的策略选择。由于博弈环境本身是透明的，研究者可以较为直接地分析参与者之间的最优响应关系，并进一步讨论纯策略或混合策略条件下的均衡结果。

与完全信息博弈相对应，不完全信息博弈是指至少存在部分与决策相关的重要信息无法被所有参与者准确掌握的情形。在这类博弈中，某一参与者可能无法明确知道对手的收益函数、策略集合、资源约束、行动能力或某些具体参数，只能依据已有经验、外部观测或先验知识形成判断。因此，参与者在进行决策时，不仅需要考虑对手可能采取何种行动，还需要考虑对手可能处于何种信息状态或具备何种能力条件。这种不确定性使得参与者的选择建立在不完全已知的信息基础之上，而不是建立在完全透明的博弈环境之中。

在形式化处理中，不完全信息博弈主要通过“类型”与“信念”加以描述。所谓类型，是指参与者所掌握但其他参与者未必知道的私有信息，它可能影响该参与者的收益结构、可选策略或行为约束。设参与者i的类型记为θ_i，则不同类型往往对应不同的收益函数或不同的行动条件。其他参与者无法直接观察到θ_i，只能依据某种先验概率分布对其进行判断。在此基础上形成的主观概率判断，即为参与者关于对手类型的信念。由此可见，在不完全信息博弈中，参与者的策略选择不再仅依赖于当前可见的博弈规则，还依赖于其对未知信息的判断。

为了将不完全信息问题转化为可分析的博弈模型，Harsanyi等人[59]提出了将参与者类型显式引入博弈结构的处理方式。在这一框架下，自然状态首先决定各参与者的具体类型，而后参与者根据自身类型及其对其他参与者类型的信念进行

决策。由此得到的博弈通常被称为贝叶斯博弈，其相应的均衡概念为贝叶斯纳什均衡。贝叶斯纳什均衡强调在给定自身类型和对其他参与者类型的信念条件下，每个参与者所采取的策略都应使其期望收益达到最大。因此，与完全信息条件下的纳什均衡相比，不完全信息博弈中的均衡分析更加强调先验判断、期望收益和信息不确定性对策略选择的影响。

对于本文所研究的黑盒模型水印问题，完全信息与不完全信息这两类情形都具有明确的建模意义。在较为理想化的分析中，可以假设防御者与攻击者都清楚彼此的策略空间、收益结构及约束条件，从而将问题建模为完全信息博弈。这样的设定有助于较为清晰地刻画攻防双方在透明条件下的基本策略关系，也便于分析不同参数变化对均衡结果的影响。然而，在实际场景中，黑盒模型水印通常具有明显的信息不对称特征。防御者往往掌握触发集构造方式、水印嵌入强度、标签映射规则以及验证阈值等私有信息，而攻击者能够获得的通常只是模型的有限外部行为信息，难以完全准确地把握水印机制的具体细节。在这种条件下，攻击者的决策不再建立在完全已知的信息基础上，而更多依赖于有限观测、经验判断和先验认知。

因此，从黑盒模型水印对抗分析的角度看，完全信息博弈与不完全信息博弈分别对应两类不同的建模视角。前者有助于揭示攻防双方在信息透明条件下的基本策略关系，后者则更适合刻画现实环境中由私有信息与有限观测所引起的决策差异。本文后续研究也将分别从这两种情形出发，对黑盒模型水印中的攻防行为进行形式化分析，从而更全面地讨论不同信息条件下的策略选择与均衡结果。

2.2.3 最优响应与纳什均衡

在博弈论分析中，最优响应与纳什均衡是描述参与者策略选择及其稳定结果的两个基本概念。前者用于刻画参与者在给定其他参与者策略条件下的最优选择，后者则用于描述所有参与者策略相互适配时所形成的稳定状态。对于本文后续讨论的黑盒模型水印攻防问题而言，这两个概念能够为分析防御者与攻击者在不同约束条件下的行为选择提供基本工具。

最优响应是指在其他参与者策略既定的条件下，某一参与者使自身收益达到

最大的策略选择。设参与者i的策略集为S_i，收益函数为$u_i(\cdot)$，其余参与者的策略组合记为s_{-i}，则参与者i的最优响应集合可表示为：

$$BR_i(s_{-i}) = \arg\max_{s_i \in S_i} u_i(s_i, s_{-i}) \tag{2.12}$$

这一表达式说明，最优响应并不一定唯一，在某些情况下，多个策略都可能使参与者获得相同的最大收益，因此最优响应通常可表示为一个集合。从分析角度看，最优响应刻画的是参与者在既定外部环境下的局部最优行为，也反映了参与者对其他各方策略的回应方式。因此，在研究博弈的均衡解之前，通常需要先明确各参与者的最优响应关系。

最优响应的意义在于，它为理解博弈中策略如何彼此制约提供了直接依据。若某一参与者改变策略，会导致其他参与者的最优响应发生变化，则博弈结果也可能随之改变。反之，若某一策略组合下各参与者都已经处于各自的最优响应状态，则该组合往往具备某种稳定性。从这一意义上说，最优响应不仅是求解均衡的基础，也有助于揭示不同参数条件下策略选择如何受到影响。在本文后续的黑盒模型水印建模中，防御者的最优响应可能体现为在给定攻击强度条件下如何选择嵌入强度或触发集参数，攻击者的最优响应则可能体现为在给定防御策略条件下如何选择攻击方式与攻击强度。

在最优响应概念基础上，可以进一步给出纳什均衡（Nash Equilibrium）的定义。设策略组合：

$$s^* = (s_1^*, s_2^*, \ldots, s_n^*) \tag{2.13}$$

为某博弈中的一个候选解，如果对任意参与者i及其任意可行策略$s_i \in S_i$，都有

$$u_i(s_i^*, s_{-i}^*) \geq u_i(s_i, s_{-i}^*) \tag{2.14}$$

则称s^*为该博弈的一个纳什均衡[60]。该定义表明，在纳什均衡下，每一位参与者当前所选择的策略，都是在其他参与者策略保持不变条件下对自身最有利的选择。因此，任何参与者都无法仅通过单方面改变自身策略而获得更高收益。正因如此，纳什均衡通常被理解为博弈中的一种稳定状态。

需要指出的是，纳什均衡未必总是以纯策略形式存在。在某些博弈中，参与者若仅能在确定性策略之间进行选择，可能不存在满足稳定条件的策略组合。为

处理这一情形，博弈论进一步引入混合策略，即允许参与者按照一定概率在多个纯策略之间进行随机化选择。纳什定理表明，对于任意有限博弈，至少存在一个混合策略纳什均衡。即使纯策略均衡不存在，博弈仍可能通过随机化策略达到某种稳定状态。本文后续研究主要关注黑盒模型水印攻防中的静态博弈模型，在具体分析中将根据问题设定讨论纯策略或混合策略条件下的均衡结果。

对于黑盒模型水印中的攻防问题而言，纳什均衡具有明确的解释意义。防御者在选择水印嵌入强度、触发集复杂度或验证阈值时，需要考虑攻击者可能采取的模型修改方式；攻击者在选择剪枝程度、微调强度或蒸馏策略时，也需要考虑防御者的水印设计方式。在此条件下，攻防双方的策略都不是孤立确定的，而是建立在对对方行为预期的基础之上。若某一组策略能够使防御者与攻击者在彼此策略既定的条件下都没有单方面改变自身选择的动机，则该组策略便可视为一个稳定的对抗结果。因此，最优响应与纳什均衡不仅为后续模型求解提供了理论依据，也为理解黑盒模型水印中不同策略组合下的结果提供了形式化工具。

总体而言，博弈论为分析黑盒模型水印中的攻防关系提供了合适的形式化工具。一方面，模型所有者与攻击者在水印嵌入、攻击实施和模型可用性保持等方面存在明显的目标冲突，双方的策略选择具有相互依赖性；另一方面，黑盒模型水印的实际部署场景又往往伴随着攻击成本约束和信息不对称特征，使得其攻防过程并非单纯的算法优化问题[61,62]，而更适合表述为参与者之间的策略选择与收益权衡问题。因此，利用博弈论中的完全信息与不完全信息分析框架、最优响应关系以及纳什均衡概念，可以较为自然地刻画黑盒模型水印中防御者与攻击者在不同条件下的策略选择及其稳定结果。这也为后续构建黑盒模型水印攻防博弈模型奠定了理论基础。

2.3 黑盒模型水印攻防建模基础

2.3.1 攻防目标与策略空间

在黑盒模型水印攻防场景中，防御者与攻击者分别面临不同的目标、约束与可选策略，其行为本质上都可归结为策略选择问题。对于防御者而言，其任务并

不只是将某种所有权标识嵌入模型，而是在模型原始任务性能、水印验证效果与水印保留能力之间实现合理平衡。对于攻击者而言，其任务也并非单纯破坏模型，而是在尽量保留模型原始功能和应用价值的前提下，削弱或移除模型中的水印信号。黑盒模型水印中的攻防关系本质上体现为双方在多重目标和有限约束下的选择问题[19]。防御者和攻击者的目标如图 2.2 所示。

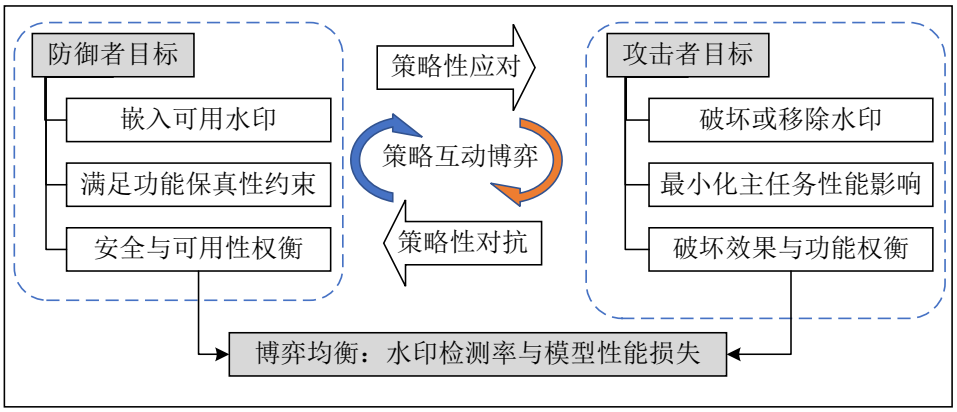

图 2.2 防御者与攻击者目标

从防御者角度来看，其主要目标是使模型在保持原始任务性能的同时，具有较高的水印验证准确率，并在模型经过一定程度修改后仍保留较好的水印可验证性。换言之，防御者希望在不显著损害模型正常功能的前提下，提高水印的稳定性和确权能力。设防御者的策略记为 s_d 其收益可理解为水印验证效果和模型原始任务性能的综合结果，同时还受到水印嵌入成本和训练约束的影响。若用 A_w 表示水印验证准确率， P 表示模型在嵌入水印后在原始任务上的计算准确率， C_d 表示防御代价，则防御者的目标可以概括为：在约束条件下尽量提高 A_w 和 P，并控制 C_d 的增加。由此可见，防御者的目标并不是单纯追求某一项指标的最大化，而是对多个指标之间的平衡进行选择。

与之相对，攻击者的目标是在尽量保持模型原始任务性能和模型使用价值的基础上，使水印验证效果尽可能下降。设攻击者的策略记为 s_a，其收益主要取决于去水印效果、攻击后模型的保留程度以及攻击成本。若以攻击后水印验证准确率的下降幅度表示去水印效果，以 P' 表示攻击后模型的原始任务性能，以 C_a 表示攻击成本，则攻击者的目标可理解为：在控制模型性能损失和攻击代价的条件下尽可能降低水印验证结果。由此可见，攻击者的选择也并非越强越好。若某种

攻击能够显著削弱水印，但同时导致模型功能严重退化或攻击代价过高，则该策略在实际意义上未必是有利的。

对于攻击者而言，其策略空间主要来源于可实施的模型修改方式及相应强度设置。若从更一般的角度看，攻击者的策略可由攻击类型、攻击强度和攻击资源分配三个方面构成。攻击类型反映攻击者选择何种方式对模型进行处理，例如参数删减、继续训练或替代模型重建；攻击强度则体现为剪枝比例、微调轮次、学习率或蒸馏规模等参数设定；资源分配则反映攻击者在有限时间、计算资源和数据资源条件下如何组织不同攻击操作。不同攻击策略对水印和模型功能的影响并不相同。较高强度的剪枝可能更容易削弱水印，但也更可能影响模型原始任务性能；更长时间的微调可能进一步降低触发响应，但也会带来更高的计算代价。

除目标和策略外，攻防双方还分别受到不同约束条件的限制。对于防御者而言，最主要的约束来自模型可用性，即水印嵌入后模型原始任务性能不能出现不可接受的退化，同时模型在正常输入上的行为不应出现明显异常。此外，防御者还受到训练成本、部署条件和验证可操作性的影响。对于攻击者而言，约束主要表现为攻击权限、可获得资源和模型可用性要求。若攻击者只能通过外部查询访问模型，则其攻击空间显然小于能够直接修改模型参数的情形；若攻击者希望保留模型的商业价值或使用价值，则任何导致模型功能显著下降的攻击策略都将受到限制。正因为双方都在不同约束条件下进行选择，黑盒模型水印中的攻防问题才具有明确的博弈分析意义。

总体来看，黑盒模型水印中的防御者与攻击者并不是在固定规则下机械执行操作，而是围绕不同目标、约束与可调变量进行策略选择。防御者试图通过合理设计触发集、控制嵌入强度并设定验证规则，提高水印的可验证性和保留能力；攻击者则试图通过选择合适的模型处理方式和攻击强度，在保留模型原始功能的同时削弱水印验证结果。正是这种在多重指标和有限约束条件下展开的选择关系，使黑盒模型水印问题能够被进一步抽象为形式化的博弈模型。

2.3.2 黑盒模型水印博弈建模

由于黑盒模型水印中的防御者与攻击者的目标存在明显冲突，且任何一方的选择都会影响另一方的收益结果，因此该问题可以自然表述为一个由参与者集

合、策略空间和收益函数组成的博弈系统。

设黑盒模型水印中的对抗过程博弈记为：

$$G=\left(N,S,U\right) \tag{2.15}$$

其中，$N=\{d,a\}$ 表示参与者集合，d 表示防御者，a 表示攻击者；$S=S_d\times S_a$ 表示策略空间，其中 S_d 和 S_a 分别表示防御者与攻击者的策略集；$U=\left(u_d,u_a\right)$ 表示收益函数集合，其中 u_d 和 u_a 分别刻画防御者与攻击者在给定策略组合下的收益。对于任意策略组合：

$$s=\left(s_d,s_a\right)\in S \tag{2.16}$$

都对应一个具体的攻防结果，包括攻击后模型的原始任务性能、水印验证效果、水印保留程度以及双方各自承担的代价。

在这一抽象框架下，防御者的策略 s_d 可以理解为对水印设计参数的选择，例如触发集构造方式、嵌入强度、验证阈值以及与训练相关的控制参数等。攻击者的策略 s_a 则可以理解为对攻击方式及攻击强度的选择，例如剪枝程度、微调参数、蒸馏规模或不同攻击方式的组合。策略一旦给定就会诱导出一个确定的模型状态和验证结果，从而进一步决定双方的收益水平。由此可见，黑盒模型水印中的攻防关系并不是孤立的技术过程，而是建立在双方策略选择上的结果映射问题。

为了量化不同策略组合所对应的结果，需要进一步定义收益函数。对于防御者而言，其收益通常与水印验证效果、水印保留能力以及模型原始任务性能保持程度正相关，而与水印嵌入代价、训练开销和部署负担负相关。因此，防御者的收益函数可以一般性地写为：

$$u_d=u_d\left(A_w,R,P,C_d\right) \tag{2.17}$$

其中，A_w 表示水印验证效果，R 表示水印保留能力或鲁棒性，P 表示模型原始任务性能，C_d 表示防御代价。相应地，攻击者的收益通常与去水印效果和模型保留程度正相关，而与攻击成本负相关，因此其收益函数可以一般性地写为：

$$u_a=u_a\left(A_w,R,P,C_a\right) \tag{2.18}$$

其中，C_a 表示攻击成本。需要指出的是，在具体建模时，收益函数不一定采用完全相同的形式，而应根据研究对象、攻击方式和分析目标作进一步设定；但从抽

象层面看，攻防双方的收益都可以理解为对验证效果、模型性能和实施代价等因素的综合度量。

在收益函数之外，约束条件同样是建模中不可忽视的组成部分。对于防御者而言，最主要的约束在于水印嵌入不能导致模型原始任务性能出现不可接受的下降；对于攻击者而言，约束则在于攻击后模型仍需保持一定的可用性，同时攻击成本不能超出其可承受范围。若分别记防御者与攻击者的约束集为Ω_d与Ω_a则有$s_d \in \Omega_d, s_a \in \Omega_a$。这说明，不同的资源条件、性能容忍范围和信息条件都会影响双方可选择的策略范围，从而进一步影响均衡结果。

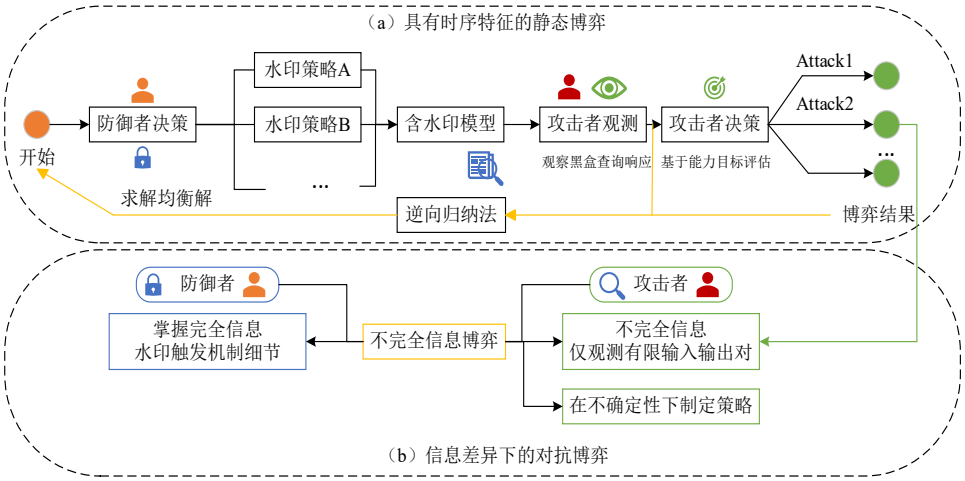

图 2.3 模型水印对抗博弈分析框架

黑盒模型水印的对抗博弈框架如图 2.3 所示，图 2.3(a)强调模型水印对抗过程中的先后顺序，图 2.3(b)则强调由私有信息和有限观测引起的信息不对称。从时序关系看，黑盒模型水印的攻防问题具有明确的先后顺序[26]。具体而言，防御者首先在训练或微调阶段完成水印嵌入，生成带水印模型；随后，攻击者在接触到该模型或其外部行为之后，再根据自身条件选择攻击策略。因此从实际背景出发，该问题具有先嵌入、后攻击的基本顺序[63,64]。不过，在后续分析中，若将双方策略选择视为在给定条件下的相互预期与相互制约，则仍可以将其表述为静态博弈框架下的策略选择问题，从而讨论双方在不同参数设定和不同信息条件下的均衡结果。

除时序特征外，黑盒模型水印攻防问题还具有明显的信息差异。防御者通常掌握完整的水印设计信息，包括触发集构造方式、嵌入参数和验证规则等；而攻

击者在多数情形下只能通过有限外部观测或模型副本推测水印机制的具体细节。这意味着，该问题既可以在理想化条件下抽象为完全信息博弈，以分析双方在透明条件下的基本策略关系，也可以进一步扩展为不完全信息博弈，以刻画攻击者在私有信息不可见条件下的选择行为[65]。

2.4 本章小结

本章围绕后续黑盒模型水印攻防博弈分析所需的技术背景与理论基础展开论述。首先，介绍了黑盒模型水印的基本框架与评价指标，说明了黑盒模型水印从触发集构造、水印嵌入到水印验证的基本实现过程，并从模型原始任务性能、水印验证效果和水印鲁棒性等方面对其评价依据进行了归纳。其次，回顾了博弈论的基本概念、完全信息与不完全信息博弈、最优响应与纳什均衡等内容。最后结合黑盒模型水印的实际部署环境，分析了对抗双方的目标与策略空间，并在此基础上对黑盒模型水印中的攻防目标、策略空间与博弈建模进行了形式化抽象。

第三章 完全信息下黑盒模型水印双人静态博弈

3.1 模型假设与问题定义

3.1.1 攻防场景与参与者设定

本章聚焦于触发集驱动的黑盒模型水印机制中防御者与攻击者之间的策略性互动问题。在当前深度学习模型广泛部署于云服务、边缘设备及第三方平台的背景下，模型所有者往往仅能通过输入输出接口与自身模型交互，而无法直接访问其内部参数或结构，即处于典型的黑盒环境。与此同时，潜在攻击者可能出于规避版权追责、转售模型或规避监管等目的，对已部署模型实施微调、剪枝、知识蒸馏等后处理操作，以削弱甚至完全移除其中嵌入的水印信号。在此现实约束下，防御者需在不显著损害模型原始任务性能的前提下，设计具备足够鲁棒性的水印机制；而攻击者则试图在可控计算成本与可接受功能退化范围内，有效破坏水印的可验证性。这一过程呈现出典型的策略依赖性与目标冲突性，天然契合博弈论的建模范式。

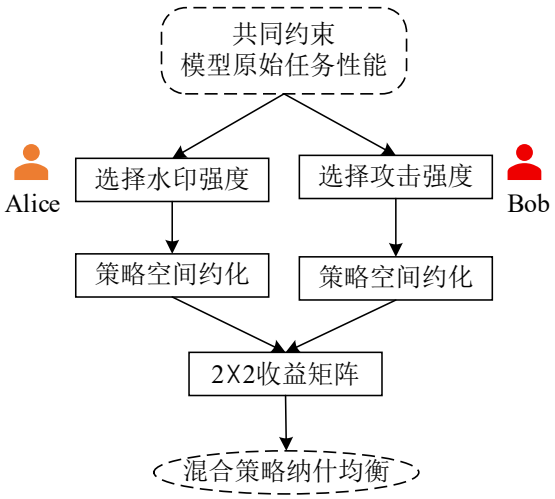

图 3.1 黑盒模型水印双人静态博弈建模流程图

为形式化该攻防场景，本文引入两名理性参与者：模型所有者记为防御者Alice，攻击方记为Bob。Alice的核心决策变量是水印嵌入强度；Bob的决策变量则是攻击强度，反映其对模型施加去水印操作的激进程度。值得注意的是，尽管

双方目标对立，但其策略选择均受到一个共同约束：模型主任务性能不得出现不可接受的退化。若水印嵌入过强或攻击过于剧烈，导致模型在目标任务上表现严重下滑，则无论水印是否存续，该模型的实用价值与商业意义都将大幅折损。因此，攻防双方实际上处于一种“竞争中带有合作约束”的复杂关系之中，这种结构性耦合为后续博弈均衡分析提供了关键前提。这一对抗过程的博弈建模流程如图 3.1 所示。

3.1.2 策略空间

设 M_0 表示仅使用正常训练样本训练所得的无水印基准模型。在触发集黑盒水印中，防御者通过在训练阶段混入一定比例的触发样本集来得到水印模型。利用触发样本集占原始训练集数据的比例刻画水印嵌入强度，并构造一组离散候选水印模型：

$$M_1, M_2, \ldots, M_N \tag{3.1}$$

其中每个 M_i 对应一种触发集比例 α_i（满足 $0=\alpha_0<\cdots<\alpha_N\le 1$），并令 $\alpha_0=0$ 对应无水印模型 M_0。因此，防御者Alice的策略空间定义为：

$$S_A=\{\alpha_0,\alpha_1,\ldots,\alpha_N\} \tag{3.2}$$

直观上，α 越大，模型越可能“记住”触发映射从而提高水印检测能力；但过大的 α 也可能引入训练分布偏移或容量占用，导致原任务泛化能力下降。因此，α 体现了水印嵌入中鲁棒性和可用性的典型权衡。

攻击者Bob的攻击强度由参数 $\beta\in[0,1]$ 表征，策略空间亦被离散化为有限集合为：

$$S_B=\{\beta_0,\beta_1,\ldots,\beta_M\} \tag{3.3}$$

其中 $0=\beta_0<\beta_1<\cdots<\beta_M\le 1$，$\beta_0=0$ 表示不实施任何攻击，$\beta_M=1$ 表示施加最大强度的去水印操作。

接下来，本文利用两个基础性能指标来刻画水印模型的原始任务性能以及其触发可检测性能。对于任意水印模型 M_i：

1） $p_i\in[0,1]$：在干净测试集上的分类准确率即原始任务性能；

2）$q_i \in [0,1]$：在触发测试集上的检测准确率即水印可验证性。

二者通常都希望尽可能高，但它们会随攻击策略改变而变化。此外，记$r_{i,j}$为水印模型M_i在遭受第j类攻击后所表现出的综合保留能力，用于概括攻击后模型在原始任务性能与水印响应两方面的稳定程度。该量作为后续收益函数中的鲁棒性参数，用于表征不同水印模型在不同攻击下的相对抗扰动能力。

3.1.3 博弈目标

本文提出以下两个核心指标来衡量策略组合(α_i, β_j)下的系统效用：

1）计算成功率（Computing Success Rate，CSR）该指标衡量系统的整体效用，综合反映了模型在不同攻防强度下的可用性与安全性：

$$\mathrm{CSR}_{i,j} = (1-\beta_j)\left[(1-\alpha_i)\,p_i + \alpha_i q_i\right] + \beta_j r_{i,j} \tag{3.4}$$

该式具有清晰的分段直觉：当攻击强度β_j较低时，模型性能主要由防御者在原始任务与水印检测率之间的内生权衡决定，即加权平均项$(1-\alpha_i)\,p_i + \alpha_i q_i$主导；而当$\beta_j$接近 1 时，系统表现则主要取决于模型在强攻击下的鲁棒性$r_{i,j}$，体现了“抗破坏能力”在极端对抗场景中的主导地位。这种分段式构造体现了如下建模思想：在弱攻击条件下，系统效用主要由模型正常使用阶段的任务性能与水印检测能力决定；而在强攻击条件下，系统能否维持基本功能更多取决于其抗破坏能力，因此以鲁棒性参数$r_{i,j}$作为主导项。计算成功率因此综合反映了模型在不同攻防强度下的可用性与安全性。

2）攻击成功率（Attack Success Rate，ASR）该指标直接量化攻击对水印及模型功能的破坏程度：

$$\mathrm{ASR}_{i,j} = 1 - r_{i,j} \tag{3.5}$$

该指标越大，表示攻击对水印及模型功能的破坏越显著，即攻击越成功。在后续收益函数中，防御者倾向于提高$\mathrm{CSR}_{i,j}$并压低$\mathrm{ASR}_{i,j}$，而攻击者则相反。

基于上述建模，本章关注的核心问题是：在触发集黑盒水印[75]场景中，面对攻击者可调的攻击强度β，防御者应如何选择触发集比例α，以在满足原任务性能约束的前提下实现尽可能稳健的水印嵌入；同时，攻击者在成本与可用性约束

下又将如何选择攻击强度与策略。因此，本章的研究目标可形式化为：在策略空间 $S_A \times S_B$ 上，结合综合性能指标 $\mathrm{CSR}_{i,j}$、鲁棒性 $r_{i,j}$ 以及后续将定义的成本收益结构，构造双方效用并求解其混合策略纳什均衡，从而得到防御者在不同攻击强度区间内对 α 的最优响应规律以及攻击强度差异与不同水印模型鲁棒性差异如何共同决定均衡结构。

3.2 收益函数设计

在触发集驱动的黑盒模型水印攻防场景中，防御者与攻击者的策略选择不仅直接影响模型的功能性与水印的可验证性，还涉及现实世界中的资源消耗、时间投入以及潜在的经济激励。若仅以模型准确率或水印检测率等纯技术指标作为决策依据，将难以刻画双方在复杂部署环境下的理性行为。为此，本文引入收益函数，在统一框架下综合反映性能表现、成本开销与战略回报，从而为后续均衡分析提供建模基础。

3.2.1 收益函数定义

借鉴对抗性机器学习与安全经济学的建模范式[66,67]，防御者（Alice）与攻击者（Bob）的总体收益表示为：

$$U_{\text{Player}} = -I_{\text{Player}} - O_{\text{Player}} + R_{\text{Player}}, \text{Player} \in \{\text{Alice}, \text{Bob}\} \tag{3.6}$$

该表达式由三部分构成：初始成本 I_{Player}、持续成本 O_{Player} 和净收益项 R_{Player}。三者共同刻画了参与者从策略准备到执行全过程中的资源投入与战略回报，体现了“成本-收益”权衡这一理性决策的基本逻辑。

初始成本 I_{Player} 表示参与者在进入博弈前必须承担的固定沉没成本[68]，其大小不依赖于具体策略选择。对防御者而言，该项涵盖模型研发、数据标注、基础训练基础设施等前期投入；对攻击者而言，则包括攻击算法开发、查询接口适配、模拟环境搭建等必要准备。尽管该成本在所有策略下保持不变，因而在最优响应分析中不影响策略排序，但其存在对于收益函数的完整性与现实映射具有重要意义，尤其在跨场景比较或长期博弈建模中不可忽略[69]。

持续成本 O_{Player} 反映参与者在实施特定策略过程中产生的可变开销[68]，其大小直接依赖于所选策略的强度。鉴于本章将防御策略抽象为触发样本比例 α_i 、攻击策略抽象为攻击强度 β_j ，为兼顾模型简洁性与现实合理性，持续成本被设定为与策略强度呈线性关系：

$$O_{i,j}^{\text{Alice}} = \alpha_i \cdot O_{\text{def}}, O_{i,j}^{\text{Bob}} = \beta_j \cdot O_{\text{att}} \tag{3.7}$$

其中 $O_{\text{def}} > 0$ 与 $O_{\text{att}} > 0$ 分别表示单位防御强度与单位攻击强度对应的边际成本系数。该设定符合直观认知：更强的水印嵌入通常需要更复杂的训练流程、更高的计算负载或更精细的触发器设计，从而带来更高的维护成本；而大规模蒸馏或深度微调等高强度攻击则往往伴随大量查询请求、长时间训练或高算力消耗，导致显著的资源支出。

收益项 R_{Player} 是收益函数的核心组成部分，用于量化策略执行后所产生的正向回报与负向惩罚。在模型水印场景中，收益不仅取决于攻防结果本身，还受到模型主任务性能退化的间接影响。这是因为，若攻击导致模型在目标任务上严重失效，即便水印被成功移除，攻击者所获得的模型亦丧失实用价值；同理，若防御者为保留水印而过度牺牲模型精度，亦会削弱其商业竞争力。因此，收益项需同时包含策略成功带来的正收益与性能退化或策略失败引发的负收益。基于此，本文将收益项分解为：

$$R_{i,j}^{\text{Player}} = -R_{\text{Player}}^{-} + R_{\text{Player}}^{+} \tag{3.8}$$

其中 R_{Player}^{-} 表征由模型功能损失或目标未达成所带来的惩罚，R_{Player}^{+} 则反映策略成功实现后的奖励。

3.2.2 防御者收益函数

结合第 3.1 节定义的计算成功率（CSR）与攻击成功率（ASR），可进一步形式化双方的收益项。对防御者Alice而言，其核心目标是在保障模型可用性的前提下最大化水印的鲁棒性。为此，其收益函数设计为：

$$U_{i,j}^{\text{Alice}} = -I_{\text{def}} - \alpha_i O_{\text{def}} - \left(\text{COO}_{i,j} + \overline{\text{CSR}_{i,j}}\right) R_{\text{def}}^{-} + \text{CSR}_{i,j} R_{\text{def}}^{+} \tag{3.9}$$

其中 $\overline{\text{CSR}_{i,j}} = 1 - \text{CSR}_{i,j}$ 表示综合成功率的互补项，反映系统整体性能的缺失程

度；$\mathrm{COO}_{i,j}$为由策略组合$\left(\alpha_i,\beta_j\right)$引起的额外性能退化项，例如主任务准确率下降超出容忍阈值的部分，可视为对“过度防御”的二次惩罚。参数$R_{\mathrm{def}}^{+}>0$与$R_{\mathrm{def}}^{-}>0$分别表示单位成功收益与单位失败惩罚的权重，体现防御者对水印可靠性与模型功能性的相对重视程度。

3.2.3 攻击者收益函数

相应地，攻击者Bob的收益函数旨在平衡攻击成本、模型可用性约束与去水印效果，其公式为：

$$U_{i,j}^{\mathrm{Bob}}=-I_{\mathrm{att}}-\beta_j O_{\mathrm{att}}-\left(\mathrm{COO}_{i,j}+\overline{\mathrm{ASR}_{i,j}}\right)R_{\mathrm{att}}^{-}+\mathrm{ASR}_{i,j}R_{\mathrm{att}}^{+} \tag{3.10}$$

其中$\overline{\mathrm{ASR}_{i,j}}=1-\mathrm{ASR}_{i,j}$反映攻击未完全成功的程度。$R_{\mathrm{att}}^{+}$为成功移除水印所带来的潜在收益；$R_{\mathrm{att}}^{-}$则量化因攻击失败或模型失效造成的资源浪费与机会成本。值得注意的是，$\mathrm{COO}_{i,j}$在双方收益函数中共享，体现了攻防双方对模型基本可用性的共同关切。该设计避免了传统零和博弈中对模型价值的忽视，更贴近实际AI服务场景中模型作为高价值资产的应用特征。

3.3 理论推导与分析

3.3.1 线性化近似

为获得可解析的均衡表达式，本文对计算成功率作局部线性化近似。该近似主要服务于理论分析，其目的在于保留性能变化趋势与策略比较关系，而非精确拟合任意水印模型在所有攻击条件下的真实响应曲线。回顾式（3.4）计算成功率的定义，加权项$\left(1-\alpha_i\right)p_i+\alpha_i q_i$表征无攻击时模型在主任务与水印检测间的综合表现。由于p_i与q_i本身依赖于α_i（例如更强的触发集注入可能提升q_i但降低p_i），该组合项通常不具备简单解析形式。然而，注意到对于任意固定的i，该表达式始终满足：

$$\min\left\{p_i,q_i\right\}\le\left(1-\alpha_i\right)p_i+\alpha_i q_i\le\max\left\{p_i,q_i\right\} \tag{3.11}$$

这表明其值域被限定在由 p_i 与 q_i 构成的区间内。基于此有界性，本文采用线性等价思想，将该组合项近似为关于 α_i 的线性函数如下：

$$(1-\alpha_i)\,p_i+\alpha_i q_i = 1-\lambda_i\alpha_i \tag{3.12}$$

其中 λ_i>0 为反映性能退化速率的斜率参数。为降低模型复杂度并突出策略差异对均衡结果的影响，本文进一步假设不同水印强度下的性能退化趋势在局部范围内近似一致，即采用统一斜率参数进行表征，即：

$$\lambda_0=\lambda_1=\ldots=\lambda_N=\lambda \tag{3.13}$$

该假设虽为理想化，但在实践中可通过控制触发样本复杂度或采用自适应嵌入策略予以逼近。在此条件下，CSR表达式简化为：

$$\mathrm{CSR}_{i,j}=(1-\beta_j)(1-\lambda\alpha_i)+\beta_j r_{i,j} \tag{3.14}$$

同时为公平刻画攻防双方对模型性能退化的共同关切，将耦合成本项 $\mathrm{COO}_{i,j}$ 设为对称线性形式：

$$\mathrm{COO}_{i,j}=k\alpha_i+k\beta_j \tag{3.15}$$

其中 k>0 为预设权重，反映单位策略强度对模型可用性的边际损害。

尽管原始策略空间 $S_A=\{\alpha_0,\ldots,\alpha_N\}$ 与 $S_B=\{\beta_0,\ldots,\beta_M\}$ 可能较大，但从理性决策角度出发，防御者倾向于选择“强且有效”的水印策略，而攻击者则偏好“弱但有效”的攻击。因此，极端无效策略如极弱水印或过度攻击在均衡中概率为零。基于此直觉，为在保留策略差异核心维度的同时获得可解析的均衡结果，本文将原始离散策略空间进一步约化为两类代表性策略：

$$S_A=\{\alpha_1,\alpha_2\},\ S_B=\{\beta_1,\beta_2\} \tag{3.16}$$

其中 α_1< α_2，β_1< β_2，分别代表“弱-强”防御与“弱-强”攻击。这一简化不仅保留了策略对比的核心维度，还使得后续均衡分析具备可解析性。

3.3.2 混合策略纳什均衡

在 2×2 博弈框架下，Alice采取的行动可以表示为 $(\Pr\{\alpha_1\},\Pr\{\alpha_2\})$，其中 $\Pr\{\alpha_1\}$ 是选择 α_1 作为防御策略的概率，且 $\Pr\{\alpha_1\}+\Pr\{\alpha_2\}=1$。类似地，Bob采取

的行动可以表示为$\left(\Pr\{\beta_1\},\Pr\{\beta_2\}\right)$，其中$\Pr\{\beta_1\}+\Pr\{\beta_2\}=1$。因此收益矩阵（Payoff Matrix）可以如表 3.1 描述。

表 3.1　2×2 博弈框架下的策略收益矩阵

策略选择概率	$\Pr\{\beta_1\}$	$\Pr\{\beta_2\}$
$\Pr\{\alpha_1\}$	$\left(U_{1,1}^{\text{Alice}},U_{1,1}^{\text{Bob}}\right)$	$\left(U_{1,2}^{\text{Alice}},U_{1,2}^{\text{Bob}}\right)$
$\Pr\{\alpha_2\}$	$\left(U_{2,1}^{\text{Alice}},U_{2,1}^{\text{Bob}}\right)$	$\left(U_{2,2}^{\text{Alice}},U_{2,2}^{\text{Bob}}\right)$

根据纳什均衡的定义，混合策略均衡需满足无偏性条件即任一方的混合策略应使其对手在所有纯策略下的期望收益相等。具体而言，对攻击者Bob，其在β_1与β_2下的期望收益应相等：

$$xU_{1,1}^{\text{Bob}}+(1-x)U_{2,1}^{\text{Bob}}=xU_{1,2}^{\text{Bob}}+(1-x)U_{2,2}^{\text{Bob}} \tag{3.17}$$

对于防御者Alice：

$$yU_{1,1}^{\text{Alice}}+(1-y)U_{1,2}^{\text{Alice}}=yU_{2,1}^{\text{Alice}}+(1-y)U_{2,2}^{\text{Alice}} \tag{3.18}$$

引入差分记号ΔU以简化表达。对Bob定义：

$$\begin{cases}\Delta U_{2,*}^{\text{Bob}}=U_{2,1}^{\text{Bob}}-U_{2,2}^{\text{Bob}}\\ \Delta U_{*,1}^{\text{Bob}}=U_{1,1}^{\text{Bob}}-U_{2,1}^{\text{Bob}}\\ \Delta U_{*,2}^{\text{Bob}}=U_{1,2}^{\text{Bob}}-U_{2,2}^{\text{Bob}}\end{cases} \tag{3.19}$$

对Alice定义：

$$\begin{cases}\Delta U_{2,*}^{\text{Alice}}=U_{2,1}^{\text{Alice}}-U_{2,2}^{\text{Alice}}\\ \Delta U_{*,1}^{\text{Alice}}=U_{1,1}^{\text{Alice}}-U_{2,1}^{\text{Alice}}\\ \Delta U_{*,2}^{\text{Alice}}=U_{1,2}^{\text{Alice}}-U_{2,2}^{\text{Alice}}\end{cases} \tag{3.20}$$

代入无偏性条件公式（3.17）和（3.18），可解得混合策略均衡：

$$x=\Pr\{\alpha_1\}=\frac{-\Delta U_{2,*}^{\text{Bob}}}{\Delta U_{*,1}^{\text{Bob}}-\Delta U_{*,2}^{\text{Bob}}},\ y=\Pr\{\beta_1\}=\frac{-\Delta U_{*,2}^{\text{Alice}}}{\Delta U_{1,*}^{\text{Alice}}-\Delta U_{2,*}^{\text{Alice}}} \tag{3.21}$$

值得注意的是，初始成本I_{def}、I_{att}在差分运算中被消去，表明均衡策略仅依赖于可变成本与收益结构，与沉没成本无关。这符合博弈论中“沉没成本不影响当前决策”的基本原理[57]。

3.3.3 均衡解释与策略退化边界

为进一步揭示均衡的可解释性，本文引入参数归一化假设：令单位攻击成本与单位防御成本分别与其对应的成功收益成比例，即：

$$O_{\text{att}} = kR_{\text{att}}^{+},\ O_{\text{def}} = kR_{\text{def}}^{+} \tag{3.22}$$

其中 k>0 为统一缩放因子。在此设定下，将第 3.2 节的效用函数代入差分表达式，并利用计算成功率与攻击成功率的线性化形式（式 3.14 与 3.5），可推导出防御者混合概率的紧凑表达：

$$\Pr(\alpha_1) = \frac{k\Delta\beta + \Delta r_{2,\star}}{\Delta r_{\star,2} - \Delta r_{\star,1}} \tag{3.23}$$

其中相关的差分项定义为：

$$\begin{cases} \Delta\beta = \beta_1 - \beta_2 \\ \Delta r_{2,\star} = r_{2,1} - r_{2,2} \\ \Delta r_{\star,1} = r_{1,1} - r_{2,1} \\ \Delta r_{\star,2} = r_{1,2} - r_{2,2} \end{cases} \tag{3.24}$$

式（3.23）具有深刻的策略含义：防御者的混合倾向由两类差异共同驱动：第一类是攻击强度差异 $\Delta\beta$，反映攻击方的行为跨度，第二类是不同水印模型在不同攻击下的鲁棒性差异 Δr，反映防御机制的结构性优势。特别地，若两种水印模型对同一攻击的鲁棒性响应高度相似（即 $\Delta r_{\star,1} \approx \Delta r_{\star,2}$），则分母趋近于零，导致混合概率发散，系统倾向于纯策略均衡。该结果也为水印算法设计提供了明确指导：不应仅追求单一水印模型的绝对鲁棒性，而应关注不同强度水印在面对多样化攻击时的相对鲁棒性差异。通过精心设计触发集或嵌入机制，可调控 Δr 的幅值，从而主动塑造均衡结构。

混合策略纳什均衡要求概率取值严格位于开区间(0,1)内。据此，对式（3.23）施加约束：

$$0 < \frac{k\Delta\beta + \Delta r_{2,\star}}{\Delta r_{\star,2} - \Delta r_{\star,1}} < 1 \tag{3.25}$$

在典型情形下例如 $\Delta r_{\star,2} > \Delta r_{\star,1}$，即强水印对弱攻击更具优势，该不等式可转化为对 $\Delta\beta$ 的区间约束：

$$\Delta\beta \in \left(-\frac{\max\left\{\Delta r_{1,\star}, \Delta r_{2,\star}\right\}}{k}, -\frac{\min\left\{\Delta r_{1,\star}, \Delta r_{2,\star}\right\}}{k} \right) \tag{3.26}$$

其中 $\Delta r_{1,\star} = r_{1,1} - r_{1,2}$，$\Delta r_{2,\star} = r_{2,1} - r_{2,2}$ 分别表示两种水印模型在攻击强度变化下的鲁棒性敏感度。

图 3.2 从几何上刻画了Alice在不同参数条件下的最优响应区域。阴影矩形表示混合策略均衡存在的可行域，即当攻击强度差异 $\Delta\beta$ 落入该区间时，Alice对强水印 α_2 、弱水印 α_1 采取概率混合。加粗线段所示的 $-\Delta_\beta$ 则给出了攻击强度差异的几何投影：当其超出阴影矩形区域边界时，混合均衡失效，最优响应退化为纯策略。右上角区域及 $1/k^2$ 标注进一步表明，成本系数 k 会改变均衡可行域的尺度与位置，因此混合策略的存在范围本质上由鲁棒性差异项与成本参数共同决定。

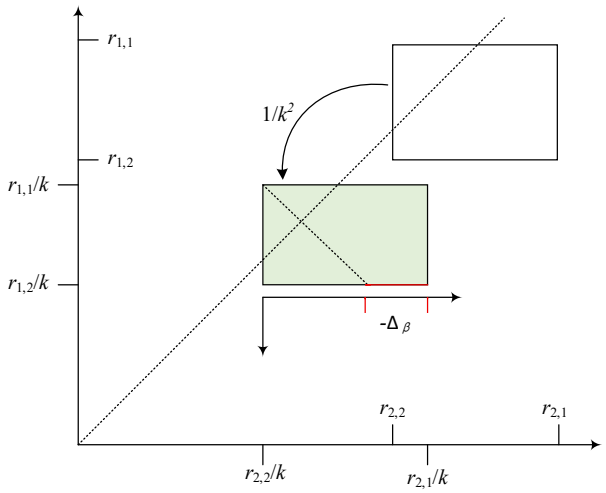

图 3.2 Alice采用混合策略下的几何域

3.4 理论结果与解释

3.4.1 博弈模型的系统性阐释

从均衡概率的显式表达式（3.23）出发出发可以看出，防御者的理性决策并不取决于孤立的水印嵌入强度或攻击强度，而取决于攻击行为变化幅度与不同水印模型鲁棒性响应差异所构成的相对结构。分子 $\left(k\Delta\beta + \Delta r_{2,\star}\right)$ 反映了攻击强度变

化对弱水印模型鲁棒性的边际影响。若攻击强度差异增大，而强水印模型对攻击变化更敏感，则防御者更倾向于降低选择弱水印的概率；反之，若弱水印在强攻击下仍能保持相对稳定，则其被选中的可能性上升。分母$\left(\Delta r_{\star,2}-\Delta r_{\star,1}\right)$则刻画了两种水印策略在面对不同攻击时的鲁棒性区分度。该值越大，表明水印强度对攻击类型的敏感性越高，从而为混合策略提供更大的存在空间。

进一步地，混合策略存在的必要条件$\Pr\{\alpha_1\}\in(0,1)$所诱导的均衡可行域式（3.26）揭示了一个关键现象：防御者的策略多样性仅在攻击行为处于中等变化区间时才得以维持。当攻击行为高度一致或攻击极端激进或保守时，均衡退化为纯策略。这意味着，在高度可预测或极端对抗的环境中，防御者缺乏实施随机化策略的动机；而在攻击模式具有一定不确定性但未达破坏性阈值的情形下，概率化防御成为理性选择。

值得强调的是，均衡条件中起决定作用的是鲁棒性指标$r_{i,j}$的差分形式，而非其绝对数值。例如，即使某一水印模型在所有攻击下均表现出较高的$r_{i,j}$，若其与另一模型的鲁棒性曲线高度平行（即$\Delta r_{\star,1}\approx\Delta r_{\star,2}$），则分母趋近于零，导致混合策略不可行。相反，若两种水印策略在面对弱攻击时表现相近，但在强攻击下分化显著（如α_2显著优于α_1），则$\left(\Delta r_{\star,2}-\Delta r_{\star,1}\right)$增大，均衡域拓宽。该结果说明，水印算法设计不应仅关注单一模型在特定攻击下的绝对鲁棒性，还应关注不同水印策略在多种攻击条件下的相对响应差异。换言之，理想的水印机制不应只能抵御特定的攻击，而应在不同攻击强度谱系上展现出可区分、可调控的鲁棒性梯度。这种梯度的存在使得防御者能够在博弈中利用策略混淆提升长期收益，同时也增加了攻击者预测与规避水印的难度。

尽管防御者与攻击者在水印验证成功率上存在根本利益冲突，但本章分析表明二者在更高层次上共享一个隐含的合作约束：模型原始任务性能的可用性。这一特征体现在收益函数中双方均包含由$\mathrm{COO}_{i,j}=k\left(\alpha_i+\beta_j\right)$引入的共同惩罚项。当任一方过度激进，不仅自身收益受损，也间接损害对方潜在利益。这是因为失效模型既无法承载有效水印，也不具备转售或再利用价值。因此该博弈呈现出典型的非零和以及部分合作属性。在均衡解中，这一特征表现为：攻击者并不总是

选择最强攻击 β_M，防御者亦不必然采用最大触发比例 α_N，而是受制于性能以及成本权衡的共同边界。

3.4.2 对水印算法设计的启示

上述理论分析为模型水印的工程实现提供了若干关键的设计启示。水印算法不应仅追求在特定攻击场景下的最大鲁棒性，而应主动构建在不同攻击强度下具有可区分响应的鲁棒性结构，例如通过多尺度触发器、自适应正则化或任务感知嵌入等方式，使不同强度的水印策略在面对轻度与重度攻击时展现出差异化的性能表现，从而扩大鲁棒性差分项的有效范围，为防御者提供更广阔的混合策略空间。同时，将水印嵌入强度例如触发样本比例或损失权重设计为连续或离散可调参数，不仅可以赋予防御者实施概率化策略的物理基础，也也增强了水印系统对未知或动态攻击的适应能力，并使防御者能够通过概率化策略提高攻击者对水印机制的预测难度[57]。此外，模型水印的长期有效性高度依赖于主任务性能的稳定性，因此在训练过程中应显式约束因水印嵌入导致的性能退化，避免以牺牲模型整体可用性为代价换取短期鲁棒性提升。最终，鉴于均衡结构对攻击行为的高度敏感性，实际部署中宜建立攻击监测与水印强度自适应调整机制，推动水印系统从静态嵌入向具备博弈意识的动态防御范式演进，从而在复杂对抗环境中实现安全性与实用性的协同优化。

3.5 本章小结

本章围绕触发集驱动的黑盒模型水印攻防问题，构建了基于博弈理论的对抗分析框架。通过形式化定义防御者与攻击者的策略空间、性能指标及收益函数结构，将水印嵌入强度与攻击强度纳入统一的决策变量体系，并引入兼顾模型原始任务性能、水印有效性与经济成本的效用函数，为双方理性行为提供了严谨的理论刻画。在此基础上，借助对计算成功率与攻击成功率的建模，结合鲁棒性参数，通过合理的线性化近似与策略空间缩减，原始高维博弈被约化为可解析的 2×2 混合策略模型，并据此推导出纳什均衡的存在条件与显式解。理论分析表明：防

御者的最优策略由攻击强度差异与水印模型间鲁棒性差分结构共同决定；混合策略均衡仅在特定参数区间内存在，其边界由鲁棒性响应的相对形态所刻画；同时，双方虽然在水印存续性上存在直接冲突，但在模型原始任务性能保持方面具有共同约束，因此该博弈呈现出一般和博弈下的非完全对立特征。研究指出相对鲁棒性结构较之单点绝对鲁棒性对均衡形态具有更强的决定作用，从而启示水印算法设计应着力优化不同防御策略在多样化攻击下的差异化响应能力，以增强系统在不确定性环境中的策略稳健性。

第四章 信息不对称下的模型水印攻防博弈分析

4.1 信息不对称博弈模型建模

在现实部署的触发集黑盒模型水印系统中，防御者与攻击者之间普遍存在显著的信息不对称性。防御者作为模型的构建者，在训练阶段主动设计并嵌入水印机制，完整掌握触发样本的构造逻辑、嵌入位置、参数配置及其在模型内部的激活路径；而攻击者通常仅能通过黑盒查询接口访问已发布的模型，无法直接观测或推断触发集的具体形式与分布特征。这种由触发信息不可观测性所引发的认知差异，使得水印攻防过程天然具备不完全信息博弈的典型属性[70]。为准确刻画此类场景下的策略互动机制，本章将触发集水印攻防建模为一个具有私有信息的不完全信息博弈。尽管从实际过程看，该问题具有“先嵌入、后攻击”的时序特征，但在分析上仍将其表述为给定信息结构下的一次性策略选择问题。这一博弈建模过程如图 4.1 所示。

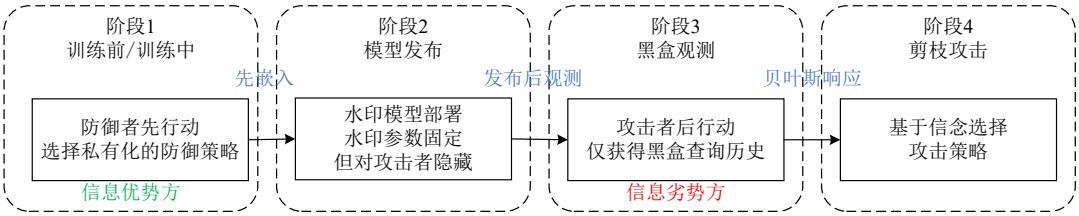

图 4.1 信息不对称下模型水印攻防博弈建模过程

4.1.1 博弈参与方与信息结构参数

本模型包含两类理性参与者：防御者（Defender）与攻击者（Attacker）。二者在目标函数与信息集上存在本质差异。防御者的目标是在维持模型主任务性能基本不变的前提下，最大化水印在对抗环境中的可验证性与鲁棒性。其决策发生在模型训练阶段，具有完全自主性与前瞻性。关键在于防御者掌握关键水印嵌入的私有信息，包括触发样本的语义结构、空间分布、激活稀疏性以及训练超参数等。这些信息构成其策略的核心组成部分，并在模型部署后对攻击者保持隐藏。而攻击者的目标则是在尽可能保留模型原始功能的前提下，通过有限的黑盒交互

削弱或消除水印响应，以规避版权验证或实现非法模型复用。由于缺乏对触发机制的先验知识，攻击者仅能基于模型输入输出行为进行逆向推断，其策略选择建立在对水印效应的不完全观测之上。换言之，攻击者面对的是一种高度受限的黑盒推断问题，既无法确认水印是否存在，也无法获知其具体形态，只能通过试探性操作评估潜在水印的存在性与敏感区域。

基于此，博弈的信息结构呈现出典型的单边私有信息特征：防御者的策略参数为其类型，属于其私有知识；攻击者虽知防御者从某一策略集合中选型，但无法确切识别其真实类型。这一设定符合贝叶斯博弈的基本范式[71]，为后续均衡分析提供了理论基础。

4.1.2 攻防者策略空间

在信息不对称博弈框架下，防御者的策略本质上体现为其在模型训练阶段对水印机制的整体设计选择。这一策略并非单一参数，而是一组相互耦合的嵌入属性，共同决定了水印在模型内部的分布形态、激活特性及其对外部扰动的敏感性。具体而言，防御者通过调节水印载体的稀疏度 ρ 来控制参与水印响应的神经元或参数子集在整个模型中的占比，较低的稀疏度有助于提升水印的隐蔽性，但可能削弱其鲁棒性；较高的稀疏度则增强水印在结构上的冗余性，提高抗攻击能力，却可能对主任务性能造成更大干扰。与此同时，防御者还需设定触发样本的复杂度 δ，用以刻画触发输入在语义、结构或数据分布层面的精细程度。复杂度越高，攻击者通过黑盒逆向推断触发模式的难度越大。此外，水印强度 γ 作为训练过程中的关键超参数，通常体现为触发样本在损失函数中的加权系数或采样频率，直接影响模型对触发输入的响应一致性与检测置信度。稀疏度、触发复杂度与水印强度这三个维度共同构成防御者的策略向量：

$$\mathbf{d} = \left[\rho, \delta, \gamma\right]^{\mathrm{T}} \tag{4.1}$$

上述参数共同决定了水印在模型中的分布特性与响应强度，并直接影响水印在面对攻击时的鲁棒性表现。需要强调的是，防御者在博弈开始前即完成水印嵌入，其策略在攻击发生时保持不变，且相关参数对攻击者不可见。

面对这一未知但固定的水印配置，攻击者的策略则体现为其在黑盒条件下实

施去水印操作的具体方式。由于无法直接定位或识别触发机制，攻击者必须依赖对模型输入输出行为的有限观测来推测潜在的水印敏感区域，并据此执行结构修改。本章则聚焦于剪枝类攻击[72]，因其在模型压缩与水印移除实践中具有广泛适用性。在此类攻击中，攻击者首先设定一个剪枝预算$k \in [0,1]$，即计划移除的模型单元比例，例如神经元、滤波器或注意力头，该剪枝预算直接决定了攻击的强度与潜在破坏范围。为在有限预算内最大化去水印效果，攻击者还需投入一定资源进行水印相关单元的估计，其估计精度通常由迭代次数或采样规模L所表征，更大的L意味着更充分的探索，但也带来更高的查询成本。然而由于黑盒反馈的噪声性与梯度估计的不稳定性，攻击过程不可避免地引入不确定性[73]，这一特性可通过探索因子ε进行建模[74]，反映策略中随机扰动或保守保留的程度。综上，攻击者的策略可形式化为向量：

$$\mathbf{a} = [k, L, \varepsilon]^{\mathrm{T}} \tag{4.2}$$

其有效性高度依赖于对防御者私有策略的间接推断能力，体现出信息劣势下的策略局限性与成本权衡。由于攻击者无法直接定位水印载体，其剪枝操作不可避免地受到估计不确定性的影响。攻击强度与攻击成本之间存在权衡关系：更大的剪枝预算或更高的估计精度通常能够更有效地抑制水印响应，但同时也可能引入更高的计算成本或模型性能损失。

4.1.3 博弈的形式化描述

综合上述要素，本章将触发集黑盒水印攻防过程建模为一个静态贝叶斯博弈，其形式化定义如下：

1）参与者集合： $N = \{\text{Defender}, \text{Attacker}\}$ ；

2）防御者类型空间：$\theta = D = \{\mathbf{d} = [\rho, \delta, \gamma]^{\top} \mid \rho \in (0,1], \delta \in \mathrm{R}_{+}, \gamma > 0\}$，其真实类型$\theta = \mathbf{d}$为私有信息；

3）攻击者信念：攻击者对防御者类型持有先验分布$P(\mathbf{d})$，通常可假设为均匀分布或基于历史经验的先验；

4）策略空间：防御者策略为类型本身即选择$\mathbf{d} \in \theta$，攻击者策略为从可观测

历史到行动的映射 $s_A : H \to A$，其中 H 为黑盒查询历史，$A = \{[k, L, \varepsilon]^{\top}\}$；

5）收益函数：双方效用 $U_D(\mathbf{d}, \mathbf{a})$ 与 $U_A(\mathbf{d}, \mathbf{a})$ 由模型主任务准确率、水印检出成功率及策略成本共同决定，具体形式将在第 4.2 节详述。

在此框架下，博弈流程为：防御者首先选择并固定其私有类型 $\mathbf{d}$；攻击者在不知 $\mathbf{d}$ 的情况下，基于先验信念与黑盒反馈选择攻击策略 $\mathbf{a}$；在本章聚焦的剪枝攻击场景下，攻击者的策略映射进一步具体化为对剪枝预算及相关估计参数的选择。最终收益由双方策略联合实现。该模型有效捕捉了实际水印攻防中“先嵌入、后攻击”的核心特征，为分析信息不对称对均衡策略与系统安全性的影响奠定了形式化基础。本章在信息不对称框架下重点分析攻击者的最优剪枝响应，并据此考察防御者私有信息对攻击效果的影响，以及攻击者如何在认知受限条件下逼近最优去水印行为。

4.2 效用函数与约束

在信息不对称博弈模型中，防御者与攻击者围绕水印鲁棒性与模型原始任务性能展开策略互动，其理性决策行为需通过可量化的效用函数加以刻画。与第三章不同，本章中攻击者对水印鲁棒性参数并不完全已知，因此效用函数中的收益预期建立在对防御者私有类型的信念之上。本节在明确性能指标体系的基础上，分别构建防御者与攻击者的效用函数结构，并引入合理的简化假设，为后续均衡求解提供解析基础。

4.2.1 性能指标与基本约束

在触发集驱动的黑盒水印[75]框架下，模型的效用价值由双重性能维度共同决定：一是主任务性能，即模型在干净输入上的预测准确率；二是水印响应性能，即模型对预设触发样本的输出一致性。记水印嵌入前模型在标准测试集上的准确率为 ACC_0，嵌入后未受攻击时的水印成功率（Watermark Success Rate，WSR）为 WSR_0。当攻击者实施剪枝操作后，这两项指标分别退化为 ACC_{post} 与 WSR_{post}。

一个实用的水印系统必须满足两个基本约束：首先，水印嵌入本身不应导致

主任务性能显著下降，即 ACC_{post} 在攻击后仍需维持在可接受阈值之上；其次，水印机制应具备足够的鲁棒性，使得 WSR_{post} 在合理攻击强度下不被完全抹除。这两个指标构成了攻防双方策略博弈的共同作用域：防御者致力于最大化 WSR_{post} 并维持 ACC_{post}，而攻击者则试图最小化 WSR_{post} 同时尽可能保留 ACC_{post}。值得注意的是，在信息不对称条件下，WSR_{post} 不仅依赖于攻击策略 $\mathbf{a}$，还隐式地受到防御者私有策略 $\mathbf{d}=[\rho,\delta,\gamma]^{\top}$ 的调制，例如高稀疏度或高触发复杂度的水印往往对盲目剪枝更具抵抗力，这一依赖关系将在后续建模中体现。

4.2.2 双方效用函数定义

防御者的根本目标是在保障模型可用性的前提下，确保水印在经受攻击后的可验证性。为此，其效用函数需综合考虑水印鲁棒性收益、原始任务性能收益以及水印嵌入带来的固有成本。给定防御策略 $\mathbf{d}$ 与攻击策略 $\mathbf{a}$，防御者效用定义为：

$$U_D(\mathbf{d},\mathbf{a})=\beta_1\cdot WSR_{\text{post}}(\mathbf{d},\mathbf{a})+\beta_2\cdot ACC_{\text{post}}(\mathbf{d},\mathbf{a})-C_D(\mathbf{d}) \tag{4.3}$$

其中 $\beta_1,\beta_2>0$ 为权重系数，反映防御者对水印验证可靠性与模型准确率的相对权重。在版权保护场景中，水印的存续通常具有优先级，故一般设定 $\beta_1\gg\beta_2$。成本项 $C_D(\mathbf{d})$ 刻画水印嵌入过程中的资源消耗，包括触发样本构造的计算开销、训练迭代次数的增加、以及因引入额外优化目标而导致的收敛不确定性。该成本是 $\mathbf{d}$ 的单调增函数：更高的水印强度 γ 或更复杂的触发模式 δ 会带来更大的训练负担与潜在性能风险。由于防御者在模型发布前已固定 $\mathbf{d}$，其效用函数主要用于刻画不同水印设计在信息不对称条件下所能达到的理论收益水平。

攻击者的目标是在维持模型基本功能的前提下，有效削弱或消除水印响应，以规避法律追责或实现非法复用。因此其效用函数需平衡模型可用性保留、水印抑制效果与攻击实施成本三重因素。在相同策略组合 $(\mathbf{d},\mathbf{a})$ 下，攻击者效用定义：

$$U_A(\mathbf{d},\mathbf{a})=ACC_{\text{post}}(\mathbf{d},\mathbf{a})-\beta_1\cdot WSR_{\text{post}}(\mathbf{d},\mathbf{a})-C_A(\mathbf{a}) \tag{4.4}$$

在该式中，第一项 ACC_{post} 表征攻击后模型的商业或实用价值；第二项则以与防御者相同的权重 β_1 惩罚残留水印，体现水印存在对攻击者的负效用；第三项

$C_A(\mathbf{a})$为攻击成本，涵盖黑盒查询开销、剪枝操作的计算资源以及因估计误差导致的冗余操作损失。由于攻击者无法直接观测到$\mathbf{d}$，其对WSR_{post}的预期具有内在的不确定性。即使采用相同的剪枝预算k，面对高隐蔽性水印时，实际去水印效果可能远低于预期。这种不确定性使得攻击者必须权衡两种互斥策略：进行激进剪枝以确保水印移除与保守操作以保全模型性能。

4.2.3 线性化假设与参数化简

为推进理论分析并获得可解析的最优策略形式，本节在保持博弈本质特征的前提下，对性能指标与成本函数引入合理近似。首先针对原始任务性能，大量实证研究表明，在适度的剪枝范围内，模型准确率的下降与剪枝比例近似呈线性关系。因此本章设定：

$$ACC_{\text{post}} \approx ACC_0 - \alpha k \tag{4.5}$$

其中α>0为准确率损失系数，其值取决于模型架构、数据集难度及剪枝粒度，可视为攻击者在特定任务下的先验知识。

其次，水印响应在结构化剪枝下通常呈现指数衰减特性[74]，尤其当水印载体分布稀疏时。结合防御者私有参数，可建模为：

$$WSR_{\text{post}}(\mathbf{d},k) \approx WSR_0 \cdot \exp\left(-\eta(\rho,\delta)\cdot k\right) \tag{4.6}$$

其中衰减率$\eta(\rho,\delta)$是稀疏度ρ与触发复杂度δ的函数：较低的ρ（意味着更集中嵌入）或较高的δ（意味着水印细节更难探测）将导致η减小，即水印对剪枝更具鲁棒性。尽管该表达式依赖于防御者私有信息，攻击者可通过历史经验或贝叶斯更新形成对η的信念分布。

最后，攻击成本主要由剪枝操作的规模决定，故近似为：

$$C_A(\mathbf{a}) \approx ck \tag{4.7}$$

其中c>0为单位剪枝成本系数，反映计算资源与时间开销。

经上述参数化处理，攻击者效用函数可简化为仅关于剪枝预算k的函数：

$$U_A(k;\eta) \approx \left(ACC_0 - \alpha k\right) - \beta_1 WSR_0 e^{-\eta k} - ck \tag{4.8}$$

该形式清晰揭示了攻击者在信息不对称条件下的优化问题：在对η存在信念分布

的前提下，选择k以最大化期望效用。这一简化不仅保留了性能-成本-鲁棒性的核心权衡结构，也为下一节推导贝叶斯最优攻击策略奠定了数学基础。

4.3 理论推导与最优解

在上一节构建的信息不对称效用框架基础上，本节致力于对攻击者在黑盒条件下的最优剪枝策略进行严格解析。通过建模剪枝操作对水印响应的抑制机制，结合性能退化与资源成本的线性近似，本节推导出攻击者效用函数关于剪枝预算k的显式表达，并在此基础上求解其最优响应策略。进一步地，通过对一阶与二阶条件的分析，系统讨论该最优解的存在性、唯一性及其对防御者私有参数的依赖结构，从而揭示信息不对称如何内生于攻防策略的均衡形成过程。

4.3.1 剪枝攻击对水印响应的影响建模

在上述效用函数与线性化近似的基础上，下面进一步刻画剪枝攻击对水印响应的影响机制。为获得可解析结果，以下分析将防御者的多维私有策略通过有效衰减因子和残余水印项进行参数化表征。在触发集水印机制中，水印信息通常被隐式编码于模型参数子空间中，其激活依赖于特定结构的触发输入。由于攻击者无法观测触发样本的具体形式或其在模型内部的映射路径，其剪枝行为本质上是一种盲性结构扰动。在此背景下，水印的鲁棒性不仅取决于嵌入强度，更关键地受制于其在参数空间中的分布特性与隐蔽程度。

设防御者通过稀疏度$\rho \in (0,1]$控制水印相关神经元或通道的比例，通过触发复杂度δ调节触发样本的语义不可预测性。攻击者通过有限次估计迭代 L以及带有随机性的选择策略ε对水印载体进行定位。在此过程中，其单位剪枝操作对水印响应的平均抑制效率可由一个有效衰减因子刻画：

$$\eta(\rho,\delta;L,\varepsilon) \in (0,1] \tag{4.9}$$

该因子在以下情形下增大：水印更集中、触发更简单或攻击者估计资源更充分。这意味着小ρ、小δ和大L下η增大，表明剪枝对水印更具破坏性。为简化符号，引入综合有效性参数：

$$a \triangleq \frac{\gamma}{\rho} \cdot \eta(\rho, \delta; L, \varepsilon) > 0 \tag{4.10}$$

其物理意义为：在给定防御配置与攻击能力下，单位剪枝预算对水印响应的指数衰减速率。值得注意的是，a 是防御者私有信息 $\mathbf{d} = [\rho, \delta, \gamma]^{\mathrm{T}}$ 与攻击者公开策略 (L, ε) 的耦合函数，构成了信息不对称博弈中“类型—行动”交互的核心桥梁。

结合水印强度 γ 用以反映训练阶段对触发样本的强化程度，水印成功率在剪枝后的残余水平可建模为：

$$WSR_{\text{post}} \leq WSR_0 \exp(-ak) + \varepsilon_{\text{res}}(\delta) \tag{4.11}$$

其中 $\varepsilon_{\text{res}}(\delta) \geq 0$ 为残余水印下界，该项仅与触发复杂度相关，在攻击强度增大时不再显著下降。这源于高复杂度触发难以被完全逆向推断，即使在强剪枝下仍可能保留微弱但可检测的响应。

4.3.2 攻击者效用函数的简化形式

基于第 4.2 节的线性化假设，原始任务准确率与攻击成本分别近似为：

$$ACC_{\text{post}} \approx ACC_0 - \alpha k, C_A(\mathbf{a}) \approx ck \tag{4.12}$$

其中 $\alpha>0$ 和 $c>0$ 为任务与资源相关的常数。将上述表达及水印衰减模型代入攻击者效用函数式（4.4）可得：

$$\begin{aligned} U_A(k; a, \delta) &\approx (ACC_0 - \alpha k) - \beta_1 \left[WSR_0 e^{-ak} + \varepsilon_{\text{res}}(\delta) \right] - ck \\ &= ACC_0 - \beta_1 \varepsilon_{\text{res}}(\delta) - \beta_1 WSR_0 e^{-ak} - (\alpha + c) k. \end{aligned} \tag{4.13}$$

由于式（4.13）中的常数项 $ACC_0 - \beta_1 \varepsilon_{\text{res}}(\delta)$ 不影响最优 k 的选择，攻击者的优化问题等价于最大化以下目标函数：

$$f(k) = \beta_1 WSR_0 \left(1 - e^{-ak}\right) - (\alpha + c) k \tag{4.14}$$

该目标函数具有清晰的经济学解释：第一项 $\beta_1 WSR_0 \left(1 - e^{-ak}\right)$ 表示因水印抑制所获得的边际收益即随 k 递增但增速递减，第二项 $(\alpha + c) k$ 为线性增长的总成本（包括性能损失与资源消耗）。该函数清晰刻画了攻击者在水印抑制收益与性能及成本损失之间的权衡关系。

4.3.3 最优剪枝预算的求解

通过对 $f(k)$ 求导，得到一阶条件：

$$f'(k)=\beta_1 WSR_0 a e^{-ak}-(\alpha+c) \tag{4.15}$$

并利用一阶条件 $f'(k)=0$，得到临界点：

$$k^*=\frac{1}{a}\ln\left(\frac{\beta_1 WSR_0 a}{\alpha+c}\right) \tag{4.16}$$

进一步计算二阶导数：

$$f''(k)=-\beta_1 WSR_0 a^2 e^{-ak}<0,\forall k\geq 0 \tag{4.17}$$

由于二阶导数 $f''(k)<0$，表明 $f(k)$ 在定义域内严格凹，故任何满足一阶条件的内点均为全局最大值。

然而，该临界点仅在满足可行性条件时有效。具体而言，当满足：

$$\beta_1 WSR_0 a\leq \alpha+c \tag{4.18}$$

有 $f'(0)\leq 0$，即效用函数在 $k=0$ 处即开始递减，此时最优策略为不实施攻击（$k^*=0$）。综合考虑剪枝预算的物理约束 $k\in[0,1]$，攻击者的最优剪枝预算最终表示为：

$$k^*=\max\left\{0,\min\left\{1,\frac{1}{a}\ln\left(\frac{\beta_1 WSR_0 a}{\alpha+c}\right)\right\}\right\} \tag{4.19}$$

该表达式揭示了攻击强度的非单调依赖性：当水印设计过于鲁棒（a 极小）或攻击成本过高（$\alpha+c$ 极大）时，理性攻击者将放弃攻击；而当水印脆弱性与攻击收益匹配良好时，攻击强度随 a 增大而先增后减，体现出典型的收益递减规律。该最优预算实质上是在攻击者基于私有类型信念所形成的期望效用下得到的最优响应。

4.3.4 数学性质分析

上述推导表明，在所设定的博弈框架下，攻击者的最优策略具有良好的数学性质：

1）存在性：由于 $f(k)$ 在紧集[0,1]上连续，根据极值定理，最大值必然存在；

2）唯一性：在 $\beta_1 WSR_0 a > \alpha + c$ 且 $k^* < 1$ 的情形下，严格凹性保证了最优解的唯一性；

3）信息敏感性：最优解 k^* 通过参数 a 间接依赖于防御者的私有策略 $\mathbf{d} = [\rho, \delta, \gamma]^\top$。这意味着即使攻击者无法直接观测 $\mathbf{d}$，其最优行为仍会因防御者的选择而系统性变化，这是贝叶斯博弈中“类型依赖策略”的典型体现。

除以上数学性质之外，残余水印项 $\varepsilon_{\text{res}}(\delta)$ 虽不影响 k^* 的取值，但会降低攻击者的绝对效用水平，从而可能影响其参与攻击的意愿，例如在扩展模型中引入攻击者进入决策。这一机制凸显了高复杂度触发设计在提升水印“不可消除性”方面的战略价值。

4.4 仿真验证与分析

为验证第 4.2 节中的关键建模假设以及 4.3 节推导的最优剪枝策略解析解，本节基于CIFAR-10 数据集与ResNet-18 模型开展系统性实验。实验设计涵盖水印嵌入性能评估、剪枝攻击行为分析、关键参数估计与理论一致性验证，以及跨种子与数据配置下的鲁棒性测试。通过定量分析原始任务准确率与水印成功率随剪枝预算k的演化规律，本文旨在实证支持信息不对称博弈框架中关于指数衰减响应、线性性能退化及最优攻击强度预测等核心假设。

4.4.1 基础性能与剪枝行为分析

实验环境基于PyTorch 1.13.0 构建，运行于NVIDIA RTX 3090 GPU上。基础模型为从头训练的ResNet-18，在CIFAR-10 上达到约 79%的初始准确率 ACC_0。水印采用BadNets[75]风格的触发机制：将一个 5×5 像素的黑色方块作为触发器，置于图像右下角（如图 4.2 所示），并将其应用于 10%的训练样本，目标类别固定为某一特定类。嵌入后，模型在干净测试集上的准确率保持在 $ACC_0 = 79.47\%$，水印成功率为 $WSR_0 = 90.39\%$，表明水印具备良好的隐蔽性与有效性。

(a)原始图像示例　　(b)触发样本示例

图 4.2 触发样本与原始样本示例

为评估模型对剪枝攻击的鲁棒性，采用基于神经元重要性估计的结构化剪枝方法[74]，剪枝预算 k 在 $\{0.005, 0.01, 0.015, 0.02, 0.03, 0.05\}$ 范围内变化，对应于 0.5% 至 5%的神经元。实验结果如图 4.3 所示：随着k增加，主任务准确率 ACC_{post} 保持基本稳定，仅从 79.47%下降至 78.85%（下降幅度 0.62%），验证了水印嵌入对原始性能影响极小；而水印成功率 WSR_{post} 则呈现单调递减趋势，从 $k=0$ 时的 90.39%下降至 $k=0.05$ 时的 85.04%，表明剪枝操作能有效抑制水印响应。

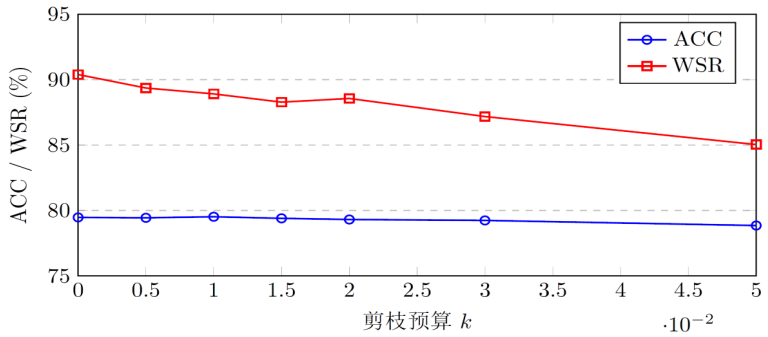

图 4.3 不同剪枝预算下含水印模型的ACC与WSR曲线图

该行为模式与第 4.2 节的线性化假设高度一致：在小剪枝预算范围内，ACC_{post} 近似为 k 的线性函数，其斜率估计值 $\alpha \approx 0.02$，即每移除 1%神经元导致约 0.02 的准确率损失，验证了式（4.5）的合理性。同时，WSR_{post} 的衰减速率显著高于模型在原始任务上准确率的下降速率，说明剪枝对水印的抑制作用远强于对原始任务的影响，这正是防御者利用信息不对称实现“低代价高收益”水印保护的关键机制。

4.4.2 参数估计与理论一致性验证

为进一步检验理论模型的预测能力，本节对不同网络层组合进行独立剪枝实验，估计各子空间的衰减率 a 与残余项 ε_{res}。如表 4.1 所示，参数 a 在不同层间存在显著差异：例如layer3.unit0.conv1 的 $a = 7.28$，表明该区域神经元对水印响应具有高度敏感性，轻微剪枝即可显著削弱水印验证准确率；而layer4.unit1.conv1 的 $a = 0.0025$，说明其对水印贡献微弱。

表 4.1 跨不同网络层组合的估计水印参数

层组合	α	a	ε_{res}	k^{*}_{theory}	k^{best}
layer4.unit0.out	0.0400	1.2677	0.0000	0.500	0.030
layer4.unit0.conv2	0.0900	0.3215	0.0000	0.500	0.030
layer4.unit0.conv1	0.0100	0.0000	0.0035	0.000	0.000
layer4.unit1.out	0.0200	0.0634	0.0000	0.500	0.030
layer4.unit1.conv2	0.0500	0.0000	0.0004	0.000	0.000
layer4.unit1.conv1	0.0100	0.0000	0.0002	0.000	0.000
layer3.unit0.out	0.1000	1.0687	0.0005	0.500	0.030
layer3.unit0.conv2	0.0900	1.3500	0.0000	0.500	0.030
layer3.unit0.conv1	0.0900	7.2813	0.0050	0.500	0.030
layer3.unit1.out	0.0100	0.0000	0.0003	0.000	0.000
layer3.unit1.conv2	0.0500	0.0075	0.0000	0.000	0.010
layer3.unit1.conv1	0.0000	0.0025	0.0000	0.000	0.000

基于全局剪枝实验数据，采用非线性回归方法拟合指数衰减模型：

$$WSR_{\text{post}}(k) = WSR_0 e^{-ak} + \varepsilon_{\text{res}} \tag{4.20}$$

得到最佳拟合参数 $a \approx 1.25$， $R^2 > 0.96$，表明模型具有极高的拟合优度。将该参数代入第 4.3 节的最优解公式（4.19）：

$$k^{*}_{\text{theory}} = \frac{1}{a} \ln\left(\frac{\beta_1 WSR_0 a}{\alpha + c} \right) \tag{4.21}$$

其中 $\beta_1 = 1$，$\alpha = 0.02$，$c = 0.01$（根据实验成本估算），计算得 $k^{*}_{\text{theory}} \approx 0.0207$。实验中通过网格搜索确定使攻击者效用最大的最优剪枝点为 $k^{*}_{\text{empirical}} = 0.0200$，二者高度吻合，验证了理论均衡分析对实际攻击行为的准确预测能力。

4.4.3 多种子与数据配置下的鲁棒性分析

为评估模型的稳定性与泛化能力，实验在五个不同随机种子下重复剪枝过程。如表 4.2 所示，尽管每次运行的剪枝路径略有差异，但拟合的衰减率a保持在 1.25 ± 0.03 范围内，且 $k=0.05$ 时的 WSR 平均值为 85.33%，标准差小于 0.5%，表明水印响应衰减模式具有高度可复现性。这些结果证实了公式（4.19）中的指数水印衰减模型在随机扰动下是稳定的，证明了修剪策略的可靠性。

表 4.2 层组合为layer4.unit0.out的多种子修剪删除结果

种子个数	a	R^2	WSR(k=0.05)(%)
0	1.229	0.967	85.04
1	1.267	0.980	84.71
2	1.293	0.979	84.77
3	1.280	0.983	84.84
4	1.199	0.975	85.33
Mean±Std	1.25±0.03	-	-

本章实验进一步比较了两种攻击场景[74]，以研究数据可用性的影响：

1）Few-shot：每类提供一个样本用于微调，此时 WSR_{post} 从 90.39%下降到 86.16%，而准确率保持在 79.23%。

2）Data-free：无真实样本，攻击者通过归一化反演重构代理输入。在这种设置中，WSR_{post} 下降至 85.22%，ACC_{post} 保持在 79.09%。

表 4.3 Few-shot和Data-free场景下基于剪枝的删除比较

场景设置	ACC_{post}	WSR_{post}
Few-shot[74]	79.23	86.16
Data-free	79.09	85.22

结果在表 4.3 中显示，实验结果说明即使在数据受限条件下，攻击者仍能有效削弱水印，但其效率略低于有监督微调场景。该现象与理论模型中残余项 ε_{res} 的存在相一致：在Data-free场景下，攻击者对触发模式的估计误差更大，导致水印残留更高，说明信息不对称会限制攻击者策略效能。

4.5 本章小结

本章围绕黑盒模型水印在信息不对称条件下的攻防过程，从博弈视角对防御者与攻击者的策略选择行为进行了系统分析。在理论分析方面，本章刻画了剪枝攻击在信息不对称条件下对水印响应的抑制机理，推导了水印成功率随攻击强度呈指数衰减的上界形式，并在此基础上求解了攻击者的最优剪枝预算。在实验方面，本章对理论模型中的关键假设与推导结论进行了验证。仿真实验结果表明，模型准确率的线性退化假设与水印成功率的指数衰减模型能够较好地刻画实际攻击行为，理论预测的最优剪枝强度与实验观测结果高度一致。同时，不同网络层剪枝效果与不同攻击信息条件下的对比分析进一步验证了水印稀疏嵌入假设及残余水印项的合理性，说明信息不对称是黑盒模型水印鲁棒性的重要来源。

第五章 黑盒模型水印静态博弈的一般框架

5.1 构建动机

在前两章中，本文分别从信息结构相对完备与显著不对称两类典型场景出发，对触发集驱动的黑盒模型水印攻防过程进行了系统性的博弈论建模与分析。第三章在假设攻防双方策略空间明确、水印机制可被完全观测的理想化设定下，构建了混合策略纳什均衡模型，揭示了防御者最优响应与攻击强度差异、水印鲁棒性结构之间的内在关联，并指出该博弈具有部分合作特征；第四章则进一步考虑现实部署环境，引入触发信息不可观测这一关键约束，将问题建模为静态贝叶斯博弈，推导出攻击者在不完全信息条件下的最优剪枝策略，并通过实验验证了信息不对称对水印鲁棒性的结构性增强作用。

表 5.1 第三章与第四章的异同关系

比较维度	第三章	第四章
信息结构	完全信息	单边私有信息
博弈类型	静态非零和博弈	静态贝叶斯博弈
防御策略	水印嵌入强度	私有水印类型
攻击策略	攻击强度	剪枝预算与估计参数
核心指标	CSR、ASR、成本	ACC、WSR、成本
均衡形式	混合策略纳什均衡	贝叶斯最优响应
主要作用	强度与鲁棒性差异	信息不对称攻击受限

前述第三章和第四章的异同关系如表 5.1 所示，尽管两章在建模范式、信息假设与求解方法上存在明显差异，但其核心逻辑具有一致性：二者均将模型水印攻防视为防御者与攻击者围绕原始任务性能、水印可验证性与策略实施成本三重目标展开的非零和博弈，且均通过形式化的策略空间、效用函数与约束条件刻画理性参与者的决策行为[12]。这种共性表明，第三章与第四章并非孤立的案例研究，而是同一类安全博弈问题在不同信息完备性条件下的具体实例：前者对应于信息较为透明的理想化情形，后者则代表典型的单边私有信息场景。

然而，现有分析亦暴露出方法论层面的局限性。第三章的均衡求解依赖于策

略空间的显式离散化与收益函数的对称可计算性，难以直接推广至高维或连续策略空间；第四章虽引入了贝叶斯框架以处理信息不对称，但其分析深度受限于特定攻击即剪枝模型与使用线性退化的简化性能假设，缺乏对更广泛攻击范式如微调、蒸馏或对抗扰动等的包容能力。更重要的是，两章分别聚焦于“信息对称”与“信息不对称”的极端情形，未能在统一框架下讨论信息对称程度变化对攻防均衡位置、策略多样性及系统鲁棒性的影响。因此有必要超越具体水印机制或攻击手段的细节，从更高抽象层次构建一个通用的模型水印攻防博弈理论框架。该框架应具备以下特征：

1）以参与者、策略超空间、信息结构与效用映射为核心要素，支持对不同水印嵌入方式如触发集、参数正则化、输出分布扰动等与攻击策略如结构修改、知识迁移、输入扰动等的统一描述；

2）显式将信息对称程度作为可调节的博弈参数[59]，从而将完全信息、不完全信息乃至部分可观测等情形纳入同一分析谱系；

3）保留前两章关键结论的可解释性，同时为动态交互、多轮学习或多方参与等复杂场景提供可扩展的理论接口。

基于此动机，本章旨在提出一个形式严谨、结构清晰且具有足够表达力的通用博弈模型，用于系统刻画黑盒模型水印攻防的本质特征。通过该框架，第三章与第四章的分析成果可被重新诠释为通用模型在特定信息配置与策略子空间下的特例，既澄清了既有结论之间的内在逻辑联系，也为未来研究更复杂的版权保护博弈如自适应水印、协同防御或多模型水印生态奠定坚实的理论基础。

5.2 通用博弈框架定义

为系统整合前文在不同信息假设下对触发集黑盒模型水印攻防问题的分析成果，并为未来研究提供可扩展的理论基础，本节提出一个通用的模型水印攻防博弈框架。该框架不依赖于特定的水印嵌入机制和攻击范式，而是通过对博弈核心要素：参与者、策略空间、效用结构与信息机制进行抽象化与参数化，旨在刻画黑盒模型水印场景中攻防互动的共性逻辑与结构性特征。这一抽象与参数化流程如图 5.1 所示。

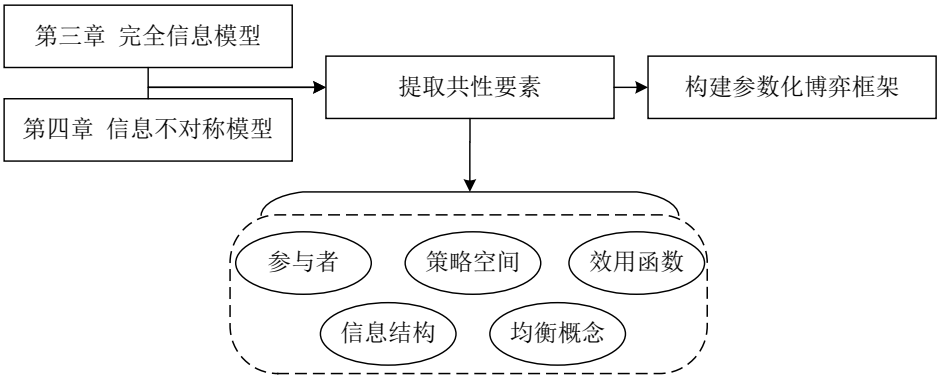

图 5.1 黑盒模型水印一般博弈框架抽象过程

5.2.1 博弈参与者与基本形式

本框架考虑一个由两名理性参与者构成的静态非零和博弈：防御者（Defender）与攻击者（Attacker）。防御者作为模型所有者，在训练阶段主动嵌入水印，其目标是在维持模型主任务性能（如分类准确率）基本不变的前提下，最大化水印在对抗环境中的可验证性与鲁棒性；攻击者作为潜在的模型滥用方，仅能通过黑盒接口访问已部署模型，其目标是在尽可能保留模型实用价值的同时，削弱、移除或规避水印响应，以实现非法复用或逃避版权追责。

在此基础上，将模型水印攻防博弈形式化为一个参数化博弈模型：

$$\mathcal{G}(\xi)=\left\langle \mathcal{P}, S_D, S_A(\xi), U_D, U_A \right\rangle \tag{5.1}$$

其中各要素定义如下：$\mathcal{P}=\{\text{Defender},\text{Attacker}\}$为参与者集合；$S_D$为防御者的策略空间，反映其在水印设计中的自由度；$S_A(\xi)$为攻击者在信息结构参数$\xi\in[0,1]$约束下的可行策略集合；$U_D$与$U_A$分别为防御者与攻击者的效用函数；$\xi$为信息对称程度参数，用于连续刻画防御者私有策略信息对攻击者的可观测水平：$\xi=1$表示完全信息即攻击者知晓全部水印细节。$\xi=0$则表示极端信息不对称即攻击者对水印机制一无所知。中间值对应不同程度的部分可观测性。该参数化形式使得不同信息条件下的具体博弈均可视为$\mathcal{G}(\xi)$在特定ξ值下的实例化，从而实现理论和实践统一。该参数并不直接进入收益函数，而是通过影响攻击者对防御者策略的可观测性、信念精度及有效策略空间而发挥作用。

信息对称程度参数ξ的引入，是本章通用框架区别于前两章具体模型的关

键。该参数并不表示攻击者是否“知道水印存在”这一状态，而是刻画攻击者对水印机制关键细节的有效掌握程度。现实中，攻击者可能通过模型架构泄露、训练数据推测、黑盒查询响应、历史攻击经验或公开论文复现等方式获得部分水印相关信息，因此完全未知与完全已知只是两个极端情形。更常见的情况是攻击者掌握有限但不完整的信息，例如知道模型可能使用触发集水印，却不知道触发样本的具体形态、目标标签、嵌入强度或敏感网络层位置。此时，ξ可以被理解为攻击者信念精度、可观测能力和有效攻击定位能力的综合表征。

从这一角度看，ξ不仅是一个形式化参数，也具有明确的安全含义。当ξ较高时，攻击者能够更准确地识别水印脆弱区域，其攻击策略更具针对性；当ξ较低时，攻击者只能采取盲性或半盲性的扰动策略，容易在水印抑制效果与模型性能损失之间产生更高的不确定性。因此，降低ξ可被视为水印防御设计中的一种结构性安全手段，其作用并不等同于单纯提高水印强度，而是通过削弱攻击者的认知能力来提高攻击代价。

5.2.2 策略空间的统一抽象

为兼容多样化的水印与攻击技术，本框架将双方策略空间定义为高维策略超空间[76]，其维度覆盖主流设计与攻击的关键自由度。防御者的策略向量$\boldsymbol{s}_D \in S_D$被抽象为四元组：

$$\boldsymbol{s}_D = (\alpha, \rho, \delta, \gamma) \tag{5.2}$$

其中：$\alpha \geq 0$用来表征模型整体对扰动的内在鲁棒性调节能力，可通过对抗训练、正则化或架构选择实现；$\rho \in (0,1]$表示水印载体在参数或神经元空间中的稀疏度，反映水印信息的集中或分散程度；$\delta \in R_+$刻画触发机制或水印信号的复杂度，包括语义自然性、分布内嵌性及逆向不可推断性；$\gamma > 0$为水印嵌入强度，通常对应于训练中触发样本的比例、损失权重或优化步长。

相应地，攻击者的策略向量$\boldsymbol{s}_A \in S_A$定义为：

$$\boldsymbol{s}_A = (\beta, k, L, \varepsilon) \tag{5.3}$$

其中：$\beta \in [0,1]$表示攻击者在性能保留与水印抑制之间的偏好权衡；$k \in [0,1]$为攻

击强度，如剪枝比例、微调步数或蒸馏温度等量化修改幅度的指标；$L \in N^{+}$反映攻击者用于水印定位或代理建模的资源投入，如查询次数、迭代轮数或采样规模；$\varepsilon \geq 0$表征策略执行中的随机性或探索性，用于建模估计误差、噪声注入或保守保留行为。

需要指出的是，策略向量中的各维度并非要求在所有场景中同时显式存在，而是为不同模型水印机制提供统一的参数接口。例如，在触发集水印中，水印强度通常表现为触发样本比例或触发损失权重，触发复杂度则对应触发图案的语义自然性、空间位置随机性或分布内嵌程度；在参数正则化水印中，水印载体稀疏度可体现为被约束参数子集的规模，嵌入强度则对应正则项权重；在输出分布水印中，复杂度可对应输出扰动模式或概率签名的不可预测性。类似地，攻击者策略中的攻击强度既可以表示剪枝比例，也可以表示微调轮数、蒸馏温度、替代数据规模或后处理扰动强度。

因此，该策略超空间的意义并不在于为所有水印方法强行规定统一的物理参数，而在于提供一种抽象坐标系，使不同水印方案和攻击方式能够在同一博弈结构中进行比较。具体研究中可根据场景选择其中若干维度进行激活，并将其余维度固定或边缘化处理。值得强调的是，上述策略空间为理论上的全集。在具体应用场景中，特定水印方案或攻击方法仅激活该超空间的子集。模型可分别视为在$\xi \approx 1$与$\xi \ll 1$条件下对S_D与$S_A(\xi)$的约束实例。其中，第三章中的水印嵌入强度和攻击强度可视为该超空间在低维离散子空间上的特例；第四章中的稀疏度、触发复杂度和剪枝预算则对应于该框架中的部分参数化实例。

5.2.3 一般形式的效用函数

为保持框架的普适性，本文不预设效用函数的具体解析形式，而仅要求其满足与原始任务性能、水印检测成功率及策略成本相关的基本单调性条件。效用函数是连接策略选择与理性决策的核心桥梁。在通用框架中，双方效用被定义为关于模型性能指标与策略成本的一般映射，避免对具体函数形式的过早限定。设$ACC \in [0,1]$为攻击后模型在干净测试集上的主任务准确率，$WSR \in [0,1]$为水印成功率，$C_D \geq 0$与$C_A \geq 0$分别为防御者与攻击者的策略实施成本如训练开销、查询

费用或计算资源。则效用函数定义为：

$$U_D = f\left(WSR, ACC, C_D\right), U_A = g\left(ACC, WSR, C_A\right) \tag{5.4}$$

其中 $f:[0,1]^2 \times R_+ \to R$ 与 $g:[0,1]^2 \times R_+ \to R$ 为连续、可微的实值函数，满足：对于防御者：$\frac{\partial f}{\partial WSR} > 0$，$\frac{\partial f}{\partial ACC} > 0$，$\frac{\partial f}{\partial C_D} < 0$ 对于攻击者：$\frac{\partial g}{\partial ACC} > 0$，$\frac{\partial g}{\partial WSR} < 0$，$\frac{\partial g}{\partial C_A} < 0$。这一抽象形式不仅涵盖了线性加权效用与指数衰减和线性成本组合，还为引入更复杂的收益结构预留了理论接口。更重要的是，它清晰揭示了攻防双方在 *ACC* 与 *WSR* 平面上的根本利益冲突：防御者追求高 *WSR* 与高 *ACC* 的联合区域，而攻击者则试图将系统推向低 *WSR* 但高 *ACC* 的象限，形成典型的非零和博弈格局。

5.2.4 信息结构参数与策略可观测性

信息结构是决定博弈均衡性质的关键变量。本框架通过参数 $\xi \in [0,1]$ 显式建模防御者策略信息对攻击者的有效可观测程度。该参数受触发隐蔽性、模型透明度、查询预算及先验知识等多重因素影响。

在形式上，ξ 通过约束攻击者的有效策略空间发挥作用：

$$S_A(\xi) = \left\{ \boldsymbol{s}_A \in S_A \mid \mathrm{E}_{\mathrm{belief}}\left[U_A\left(\boldsymbol{s}_D, \boldsymbol{s}_A\right) \right] \right\} \tag{5.5}$$

当 ξ 较高时：攻击者可构建对 s_D 的高精度信念分布，从而在接近完整策略空间 S_A 中优化；当 ξ 较低时：其信念高度不确定，导致部分高风险策略因预期效用过低而被排除，即 $S_A(\xi)$ 收缩。这种依赖关系使得信息不对称不再仅是背景假设，而是内生于博弈结构的动态调节变量。本节所提出的通用博弈框架 $\mathcal{G}(\xi)$ 通过抽象化参与者、策略、效用与信息四大要素，实现了对模型水印攻防问题的统一描述。它既保留了具体模型的解释力，又具备足够的灵活性以容纳更广泛的水印与攻击范式，为后续开展跨场景比较、均衡泛化性分析及动态扩展研究提供了严谨而通用的理论平台。

5.3 均衡条件与推论

5.3.1 均衡概念与存在性条件

在第 5.2 节所构建的通用模型水印攻防博弈框架$\mathcal{G}(\xi)$基础上，本节致力于分析该博弈的均衡性质，探讨纳什均衡的存在性、形式多样性及其对信息结构参数ξ的依赖关系。由于防御者与攻击者分别从策略空间S_D与$S_A(\xi)$中选择策略，以最大化各自效用U_D与U_A，且双方在水印存续性上存在直接利益冲突，而在模型主任务性能保持方面又具有共同约束，因此该问题属于典型的非零和博弈。

在通用博弈$\mathcal{G}(\xi)$中，防御者与攻击者分别从其可行策略集S_D与$S_A(\xi)$中选择策略$\boldsymbol{s}_D$和$\boldsymbol{s}_A$，以最大化各自的效用函数$U_D(\boldsymbol{s}_D,\boldsymbol{s}_A)$与$U_A(\boldsymbol{s}_D,\boldsymbol{s}_A)$。由于双方目标在水印存续性上存在根本冲突，且效用函数非完全对立。因其均依赖于主任务性能ACC，该博弈属于非零和博弈。在此类博弈中，纳什均衡作为策略组合的稳定性标准，自然成为刻画理性互动结果的重要概念。

为确保均衡的存在性，引入以下标准假设：

1）策略紧致性：对任意$\xi\in[0,1]$，S_D与$S_A(\xi)$均为非空、有界且闭的子集。

2）效用连续性与拟凹性：U_D关于$\boldsymbol{s}_D$连续且拟凹，U_A关于$\boldsymbol{s}_A$连续且拟凹。

3）信息结构兼容性：参数ξ仅通过约束$S_A(\xi)$或改变攻击者对U_A的期望形式影响博弈，而不破坏效用函数的连续性或策略空间的拓扑性质。

在上述条件下，根据Glicksberg对非零和博弈的推广结果[71]，$\mathcal{G}(\xi)$在混合策略空间中至少存在一个纳什均衡。这一结论说明，只要具体水印机制与攻击方式能够纳入本章提出的策略与效用结构，攻防互动在理论上就可被刻画为某种稳定的策略组合。换言之，前文第三章与第四章中所讨论的具体均衡结果，都可以视为该一般框架在特定信息结构与策略约束下的实例化表现。

5.3.2 纯策略与混合策略均衡的讨论

尽管混合策略均衡在一般条件下必然存在，但在特定参数配置下，博弈也可能支持纯策略纳什均衡。例如，当信息对称程度较高即$\xi\to 1$，攻击者能够准确

估计防御者的水印参数，其最优响应往往集中于某一确定性攻击强度k^*，而防御者亦可据此选择最优嵌入策略，形成双方均无单方面偏离动机的纯策略组合。

然而，在更贴近现实的信息不对称场景即$\xi \ll 1$中，由于攻击者对防御者私有类型存在不确定性，其最优行为通常表现为对多种攻击强度的概率化选择，以平衡“过度攻击导致性能崩溃”与“攻击不足导致水印残留”的风险。此时纯策略均衡可能不存在，而混合策略均衡成为描述稳定状态的唯一合理形式。值得指出的是，前文第四章中攻击者最优剪枝预算的解析表达，并不意味着混合策略在不完全信息场景下失效，而是因为在连续策略空间与严格凹效用条件下，均衡可能退化为集中于单一点的概率分布。由此可见，纯策略均衡与混合策略均衡并非彼此割裂，而可视为通用框架在不同策略结构与信息条件下的不同表现形式。

5.3.3 信息对称性参数对均衡的影响

信息结构参数ξ作为本框架的核心调节变量，对均衡位置具有决定性影响。形式上，均衡策略组合$\left(\boldsymbol{s}_D^*(\xi), \boldsymbol{s}_A^*(\xi)\right)$可视为$\xi$的函数。一般而言，随着$\xi$的减小，均衡解呈现出系统性的偏移趋势：当$\xi$较高时，攻击者具备更强的水印识别与定位能力，其有效策略空间更接近完整空间，因而能够更有针对性地选择攻击方式。在这种情况下，均衡结果往往更有利于攻击者，防御者通过信息隐藏所获得的结构性优势相对减弱。此时均衡中的WSR_{post}较低，攻击者效用较高，博弈格局偏向攻击方。而当ξ较低时，攻击者难以准确区分与水印相关和无关的模型结构，其策略选择更依赖有限观测与保守判断。这会降低攻击资源投入的边际收益，并使部分高风险、高强度的攻击策略因预期效用不足而失去吸引力。相应地，均衡结果通常会向更有利于防御者的方向偏移，表现为较高的水印保留能力或较低的攻击有效性。

因此，在合理的效用单调性与策略空间收缩条件下，在一般的模型水印对抗博弈中，防御者的均衡效用$U_D^*(\xi)$是ξ的非增函数，而攻击者的均衡效用$U_A^*(\xi)$是ξ的非减函数。换言之，信息不对称性本身构成一种内生的防御机制，其强度直接决定了攻防力量的对比格局。这一结论从博弈角度说明，信息不对称不仅是模型水印场景中的客观约束，也可被理解为一种具有安全增益的结构性资源。

推论 5.1：信息不对称的防御增益。在效用函数满足基本单调性条件，且攻击者有效策略空间随信息对称程度ξ的增大而单调扩张的条件下，攻击者均衡效用关于ξ非减，防御者均衡效用关于ξ非增。换言之，信息对称程度越高，攻击者越能够利用更精确的水印认知制定有效攻击策略；信息对称程度越低，防御者越能通过隐藏水印结构获得额外防御收益。

该推论表明，信息隐藏并非仅是黑盒水印场景中的被动约束，而可以被视为一种主动安全资源。当攻击者难以准确识别触发样本结构、嵌入位置或水印载体分布时，其可行攻击策略虽然形式上仍然存在，但有效性会因信念误差和定位不确定性而下降。因此，降低ξ可以在不显著提高水印嵌入强度的情况下提升攻击者的预期攻击代价，从而改善防御者在均衡中的收益位置。

5.3.4 防御优势、攻击代价与信息隐藏替代效应

基于均衡分析，可进一步刻画防御优势与攻击代价之间的非线性关系。定义防御优势为均衡状态下水印成功率$WSR^{*}_{\text{post}}(\xi)$或防御者效用$U^{*}_{D}(\xi)$，攻击代价为达到特定水印抑制水平所需的性能损失$\Delta ACC$或资源消耗$C_A$。在上述分析基础上，还可以进一步讨论防御优势与攻击代价之间的关系。这里所说的防御优势可以由均衡状态下的水印成功率、鲁棒性水平或防御者效用加以刻画；攻击代价可由达到特定去水印目标所需承担的性能损失、资源消耗或策略实施成本来表征。

从一般框架出发，当信息对称程度降低时，攻击者为达到相同的水印抑制目标，通常需要投入更多估计资源、采取更激进的攻击策略，或者承受更大的模型性能退化风险。也就是说，信息不对称往往会提高攻击者实现同等攻击效果的代价，并降低其攻击收益的确定性。

与此同时，防御者并不一定需要单纯通过增大水印强度来提升系统安全性。相较于直接增强嵌入幅度，采用触发机制复杂化、载体分散化、参数冗余设计等方式扩大攻防双方之间的信息鸿沟，往往更有可能在较低额外成本下提升攻击难度。这表明，模型水印的有效设计不仅取决于鲁棒性绝对值本身，还取决于其在博弈结构中对攻击者策略空间和攻击成本的影响方式。

因此，从通用框架视角看，理想的水印机制应不仅追求在单一攻击条件下保

持较高性能，更应关注如何通过结构设计提高攻击者的决策不确定性和实施代价。该认识为理解模型水印中的“信息隐藏优势”提供了更一般的理论解释，也为后续开展动态水印、自适应防御和多模型协同保护研究提供了分析基础。

推论 5.2：水印强度与信息隐藏的替代关系。在攻击者成本函数随攻击强度递增，且攻击效果依赖于其对水印载体定位能力的条件下，防御者可以通过提高水印嵌入强度或降低信息对称程度两种路径提升均衡防御收益。前者直接增强水印响应的稳定性，后者则通过增加攻击者的识别难度与定位误差提高攻击代价。

该推论说明，水印鲁棒性并不只来源于更强的嵌入强度。若单纯提高水印强度，可能会引入原始任务性能下降、训练成本增加或水印可检测性增强等副作用；而通过提升触发样本复杂度、分散水印载体位置、引入响应随机化或增强水印机制的不可推断性，则可以在较低性能损失下削弱攻击者的有效策略空间。因此，在实际水印设计中，应在“强嵌入”和“难识别”之间取得平衡，使水印机制同时具备鲁棒性与隐蔽性。

5.4 框架扩展

第 5.2 节所提出的通用模型水印攻防博弈框架$\mathcal{G}(\xi)$在静态、双参与者、非合作设定下成功统一了前文对不同信息结构下水印攻防行为的分析。然而现实中的模型版权保护场景往往呈现出更高的复杂性：攻防交互可能跨越多个时间阶段，参与主体可能不止一方，水印机制本身可能涉及内部组件协同，且模型形态已从传统判别式网络扩展至生成式与多模态系统。本节所讨论的扩展并非对前文静态框架的简单外推，而是说明该框架具有模块化特征：参与者集合可以扩展，策略空间可以随时间变化，效用函数可以替换为不同任务指标，信息结构参数也可以由静态常数扩展为动态演化变量。因此，通用框架既可作为当前黑盒水印攻防问题的总结模型，也可作为后续研究复杂版权保护场景的基础模板。

5.4.1 动态博弈与重复博弈视角

在实际部署中，模型水印的生命周期通常包含嵌入、发布、攻击、检测与可

能的更新等多个环节，攻防双方的行为具有明显的时序依赖性与策略演化特征。例如，防御者可能在检测到水印被削弱后，通过重新训练或微调嵌入新水印；攻击者则可能基于历史攻击效果如水印残留率、性能损失等动态调整后续攻击强度或资源分配。此类场景天然契合动态博弈或重复博弈的建模范式。在通用框架下，可将单阶段博弈$\mathcal{G}(\xi)$扩展为T阶段的序列博弈$\{\mathcal{G}_t(\xi_t)\}_{t=1}^{T}$，其中每阶段的策略空间$S_D^t, S_A^t(\xi_t)$与信息参数$\xi_t$可随时间演化。参与者的总效用由跨阶段的折现累计收益定义：

$$U_D^{\text{total}} = \sum_{t=1}^{T} \delta^{t-1} U_D^t\left(\mathrm{s}_D^t, \mathrm{s}_A^t\right), U_A^{\text{total}} = \sum_{t=1}^{T} \delta^{t-1} U_A^t\left(\mathrm{s}_D^t, \mathrm{s}_A^t\right) \tag{5.6}$$

其中$\delta \in (0,1]$为折扣因子，反映对未来收益的重视程度。在此设定下，策略承诺、信誉积累以及学习效应均可纳入均衡分析。尽管求解此类动态均衡通常涉及复杂的递归方法，但通用框架为其提供了清晰的建模接口，使得本框架不仅能描述一次性攻防，还能刻画长期对抗中的策略适应与演化规律。

5.4.2 多参与者博弈扩展

除了一对一的攻防关系，现实水印生态系统常涉及多方互动。典型情形包括单一模型面临多个独立或协同的攻击者或是模型由多个利益相关方共同开发与维护，因此需要协调各方的水印策略，例如联邦学习场景中的多方协同模型[77]。存在第三方验证者或监管机构介入水印确权过程[78,79]。对此，通用框架可通过扩展参与者集合$\mathcal{P}$自然推广至多方非零和博弈。设$\mathcal{P} = \{D_1, \ldots, D_m\} \cup \{A_1, \ldots, A_n\}$，其中$D_i$为防御者，$A_j$为攻击者。各参与者的策略空间与效用函数可分别定义为S_{D_i}, U_{D_i}与$S_{A_j}(\xi_{ij}), U_{A_j}$，其中$\xi_{ij}$用于表征攻击者$A_j$对防御者$D_i$策略的信息可观测程度。在此扩展下，攻击者之间可能存在竞争或协作；防御者之间则可能通过策略协调或水印正交化提升整体鲁棒性。核心冲突即水印存续性与模型可用性的权衡依然保持不变。

5.4.3 合作博弈视角下的防御结构分析

在非合作博弈主线之外，通用框架还可引入合作博弈论（cooperative game

theory）工具，用于分析防御者内部水印机制的结构性贡献。具体而言，可将水印载体的不同组成部分，例如特定网络层、通道组或触发特征子集视为合作博弈中的“虚拟参与者”，其联盟收益定义为该子集对整体水印鲁棒性的边际贡献。形式上，设 $N=\{1,2,\ldots,n\}$ 为水印组件索引集，定义特征函数 $v:2^N\to R_+$，其中 $v(S)$ 表示子集 $S\subseteq N$ 单独存在时所能维持的水印成功率（或等效效用）。在此基础上，Shapley值 $\phi_i(v)$ 可用于量化组件 i 的平均边际贡献[80]：

$$\phi_i(v)=\sum_{S\subseteq N\{i\}}\frac{|S|!(n-|S|-1)!}{n!}\left[v(S\cup\{i\})-v(S)\right] \tag{5.7}$$

该分析为理解水印设计机理提供了微观视角。例如，若Shapley值高度集中于少数组件，则表明水印脆弱性高；反之，若贡献分布均匀，则说明水印具有良好的冗余性与分散性。合作博弈在此与主框架中的非合作博弈形成内外互补。

5.4.4 向生成模型与多模态水印的泛化

随着深度学习模型形态的演进，水印研究已扩展至生成模型和多模态系统，如GANs[81]、扩散模型[82]、CLIP[83]和Flamingo[84]。通用博弈框架凭借其抽象性与模块化设计，可为此类场景提供可扩展的建模思路：

1）策略空间扩展：容纳生成模型特有的控制变量如潜空间扰动幅度、采样步长调节或多模态对齐参数。

2）效用函数重定义：以FID[85]、CLIP Score[86]或生成多样性替代分类准确率，同时保留水印可验证性作为核心目标。

3）信息结构参数ξ：仍有效刻画攻击者对生成水印机制的认知程度。

例如，在文本到图像生成模型中，防御者策略 $\boldsymbol{s}_D$ 可包含触发词复杂度 δ 与潜空间扰动强度 γ，攻击者策略 $\boldsymbol{s}_A$ 可包含微调数据量 L 与生成修复强度 k，其基本分析思路可在 $\mathcal{G}(\xi)$ 框架下进行适配。

5.5 本章小结

本章在前文对触发集黑盒模型水印攻防问题分别于本文第三章中信息相对

完备与第四章中显著不对称条件下进行建模与分析的基础上，提出并构建了一个更具一般性的模型水印攻防博弈理论框架。该框架不局限于具体水印机制或攻击算法，能够在统一形式下描述不同信息结构、策略空间与效用目标下的攻防关系，为后续研究更复杂的模型水印攻防问题提供了可扩展的理论参考。

第六章 总结与展望

6.1 总结

随着深度神经网络在工业、医疗、金融等关键领域的深度部署，深度模型已成为承载高价值知识产权的重要数字资产。如何在保持模型原始任务性能的前提下，实现模型版权归属的可验证性，已成为人工智能安全与可信计算领域的重要研究课题。黑盒模型水印因无需访问模型内部结构、兼容现有部署流程且具备良好隐蔽性，被视为一种具有应用前景的版权保护方法。然而，现有模型水印研究主要围绕方法设计展开，对模型所有者与攻击者在不同约束条件下的行为选择、收益权衡及均衡关系仍缺少系统的理论刻画，因而难以为模型水印的对抗分析与方法设计提供充分的理论依据。

针对上述挑战，本文以博弈论为理论分析工具，对基于触发集的黑盒模型水印的对抗问题进行了系统性研究，形成了从具体博弈建模到一般框架抽象的分析思路。全文首先梳理了黑盒模型水印技术与博弈论的基础理论，明确将模型水印对抗过程刻画为防御者（模型所有者）与攻击者（模型窃取方）围绕网络模型原始任务性能、水印成功率与策略实施成本三重目标展开的非零和博弈。在此基础上，本文依次研究完全信息条件下的黑盒模型水印双人静态博弈、信息不对称条件下的水印攻防博弈以及黑盒模型水印静态博弈的一般框架。本文的理论贡献主要体现在以下三个方面：

1）构建了黑盒模型水印的对抗博弈模型。本文将水印嵌入强度与攻击强度作为攻防双方的策略变量，结合模型原始任务性能、水印成功率和策略实施成本构建收益函数，并将该博弈约化为可解析的混合策略模型。本文求解了混合策略纳什均衡，分析了攻击强度差异与水印鲁棒性差分结构对均衡结果的影响。

2）分析了信息不对称条件下水印攻防均衡的变化规律。针对攻击者难以获知触发集细节和水印嵌入方式的实际情况，本文以剪枝攻击为典型场景，分析了攻击者在有限信息条件下的策略选择，并求解了其在模型性能保持、攻击成本与水印抑制收益之间权衡下的最优剪枝预算，结合实验验证了理论分析的有效性。

3）提出了黑盒模型水印静态博弈的一般框架。在具体博弈模型分析的基础上，抽象参与者、策略空间、效用函数与信息结构等关键要素，构建了参数化的黑盒模型水印攻防博弈框架。该框架通过信息对称程度参数统一描述不同信息条件下的对抗关系，为后续复杂模型水印攻防问题研究提供了理论参考。

6.2 展望

尽管本文围绕黑盒模型水印攻防问题开展了博弈建模与均衡分析，并构建了一般框架，但受研究对象、信息假设和求解方法等因素限制，仍存在进一步拓展的空间。随着深度学习模型部署场景的不断复杂化以及去水印攻击手段的持续演化，模型水印攻防问题将呈现出更强的动态性、多主体性和跨模型形态特征。未来研究可从以下几个方面继续展开：

1）本文主要聚焦于单阶段静态博弈，假设攻防双方在给定信息结构下完成一次性策略选择。然而，在实际应用中，模型水印通常经历嵌入、发布、攻击、检测和更新等多个阶段。未来可在本文一般框架的基础上引入动态博弈或重复博弈模型，将水印攻防过程扩展为多阶段序列决策问题[64]，进一步分析防御者如何根据攻击反馈调整水印策略、攻击者如何利用历史观测更新攻击方案，以及策略承诺和长期收益等因素对均衡结果的影响。

2）本文以单一防御者和单一攻击者为基本设定，而实际模型水印生态中往往存在多个利益相关方。例如，同一模型可能同时面临多个攻击者的独立攻击或协同攻击；在联邦学习、模型协作开发等场景中，也可能存在多个模型所有者共同维护和验证水印的情况[87]。未来可将参与者集合扩展至多防御者或多攻击者情形，分析多主体策略互动对水印鲁棒性、攻击成本和收益分配的影响。同时，也可引入合作博弈工具，研究多防御者场景下的贡献度评估与收益分配问题。

3）本文主要围绕判别式分类模型展开分析，而当前人工智能模型形态正逐步扩展至扩散模型与大语言模型等多模态系统[85,86]。未来可结合生成模型和多模态模型的任务特征，对本文框架中的策略空间、效用函数和评价指标进行适配。例如，在生成模型场景中，可将生成质量、语义一致性和水印可验证性共同纳入

效用函数；在多模态场景中，可进一步考虑跨模态触发、输出一致性和语义对齐程度等因素，从而提升博弈框架对新型模型水印问题的适用性。

4）本文主要通过解析推导获得均衡结果，但当策略空间高维、效用函数非线性或攻防反馈复杂时，闭式均衡解往往难以获得。未来可结合强化学习与演化博弈等计算方法，对复杂水印攻防场景中的近似均衡进行求解。例如，可将防御者的水印嵌入参数视为可优化变量，通过迭代训练获得更具适应性的防御策略。此类研究有助于弥补解析模型在复杂场景下的表达和求解限制，并为模型水印系统的实际部署提供更具操作性的策略参考。

参考文献

[1] Chai J, Zeng H, Li A, et al. Deep learning in computer vision: A critical review of emerging techniques and application scenarios[J]. Machine Learning with Applications, 2021, 6: 100134.

[2] Ahmed N, Saha A K, Al Noman M A, et al. Deep learning-based natural language processing in human–agent interaction: Applications, advancements and challenges[J]. Natural Language Processing Journal, 2024, 9: 100112.

[3] Chib P S, Singh P. Recent advancements in end-to-end autonomous driving using deep learning: A survey[J]. IEEE Transactions on Intelligent Vehicles, 2023, 9(1): 103-118.

[4] Wainberg M, Merico D, Delong A, et al. Deep learning in biomedicine[J]. Nature Biotechnology, 2018, 36(9): 829-838.

[5] Ozbayoglu A M, Gudelek M U, Sezer O B. Deep learning for financial applications: A survey[J]. Applied Soft Computing, 2020, 93: 106384.

[6] Jin Biao, Lin Xiang, Xiong Jinbo, et al. Intellectual property protection of deep neural network models based on watermarking technology[J]. Journal of Computer Research and Development, 2024, 61(10): 2587-2606.

[7] Shafran A, Shumailov I, Erdogdu MA, et al. Beyond Labeling Oracles: What does it mean to steal ML models?[J]. arXiv preprint arXiv:2310.01959, 2023.

[8] Zhao K, Li L, Ding K, et al. A systematic survey of model extraction attacks and defenses: State-of-the-art and perspectives[J]. arXiv preprint arXiv:2508.15031, 2025.

[9] Mosafi I, David E, Altshuler Y, et al. DNN intellectual property extraction using composite data[J]. Entropy, 2022, 24(3): 349.

[10] Li P, Huang J, Zhang S. LicenseNet: Proactively safeguarding intellectual property of AI models through model license[J]. Journal of Systems Architecture, 2025, 159: 103330.

[11] 吴汉舟, 张杰, 李越, 等. 人工智能模型水印研究进展[J]. 中国图象图形学报, 2023, 28(6): 1792-1810.

[12] Li Y, Wang H, Barni M. A survey of deep neural network watermarking techniques[J]. Neurocomputing, 2021, 461: 171-193.

[13] Li Q, Xu G. Deep neural network fingerprinting by general examples[J]. Neurocomputing, 2025: 132184.

[14] Zhong Q, Zhang L Y, Zhang J, et al. Protecting IP of deep neural networks with watermarking: A new label helps[C]//Pacific-Asia Conference on Knowledge Discovery and Data Mining. Cham: Springer International Publishing, 2020: 462-474.

[15] Zhang J, Gu Z, Jang J, et al. Protecting intellectual property of deep neural networks with watermarking[C]// Proceedings of the 2018 ACM Asia Conference on Computer and Communications Security. 2018: 159-172.

[16] Boenisch F. A Systematic Review on Model Watermarking for Neural Networks[J]. Frontiers in Big Data, 2021, 4: 729663.

[17] 郭钊均, 李美玲, 周杨铭, 等. 人工智能生成内容模型的数字水印技术研究进展[J]. 网络空间安全科学学报, 2024, 2(1): 13-39.

[18] 刘安安, 苏育挺, 王岚君, 等. AIGC 视觉内容生成与溯源研究进展[J]. 中国图象图形学报, 2024, 29(6): 1535-1554.

[19] Lukas N, Jiang E, Li X, et al. Sok: How robust is image classification deep neural network watermarking?[C]//2022 IEEE Symposium on Security and Privacy (SP). IEEE, 2022: 787-804.

[20] Uchida Y, Nagai Y, Sakazawa S, et al. Embedding watermarks into deep neural networks[C]//Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval. 2017: 269-277.

[21] Darvish Rouhani B, Chen H, Koushanfar F. Deepsigns: An end-to-end watermarking framework for ownership protection of deep neural networks[C]//Proceedings of the 24th International Conference on Architectural

Support for Programming Languages and Operating Systems. 2019: 485-497.

[22] Liu H, Weng Z, Zhu Y. Watermarking Deep Neural Networks with Greedy Residuals[C]//Proceedings of the 38th International Conference on Machine Learning. Virtual Event: Proceedings of Machine Learning Research, 2021, 139: 6978-6988.

[23] Wang T, Kerschbaum F. Riga: Covert and robust white-box watermarking of deep neural networks[C]// Proceedings of The Web Conference 2021. 2021: 993-1004.

[24] Fan L, Ng K W, Chan C S, et al. Deepipr: Deep neural network ownership verification with passports[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2021, 44(10): 6122-6139.

[25] Zhang J, Chen D, Liao J, et al. Passport-aware normalization for deep model protection[J]. Advances in Neural Information Processing Systems, 2020, 33: 22619-22628.

[26] Lou X, Guo S, Li J, et al. Ownership verification of DNN architectures via hardware cache side channels[J]. IEEE Transactions on Circuits and Systems for Video Technology, 2022, 32(11): 8078-8093.

[27] Adi Y, Baum C, Cisse M, et al. Turning your weakness into a strength: Watermarking deep neural networks by backdooring[C]//27th USENIX Security Symposium (USENIX Security 18). 2018: 1615-1631.

[28] Guo J, Potkonjak M. Watermarking deep neural networks for embedded systems[C]//Proceedings of the International Conference on Computer-aided Design. 2018: 1-8.

[29] Guo J, Potkonjak M. Evolutionary trigger set generation for dnn black-box watermarking[J]. arXiv preprint arXiv:1906.04411, 2019.

[30] Le Merrer E, Perez P, Trédan G. Adversarial frontier stitching for remote neural network watermarking[J]. Neural Computing and Applications, 2020, 32(13): 9233-9244.

[31] Chen H, Rouhani B D, Koushanfar F. Blackmarks: Blackbox multibit

watermarking for deep neural networks[J]. arXiv preprint arXiv:1904.00344, 2019.

[32] Namba R, Sakuma J. Robust watermarking of neural network with exponential weighting[C]//Proceedings of the 2019 ACM Asia Conference on Computer and Communications Security. 2019: 228-240.

[33] He Z, Zhang T, Lee R. Sensitive-sample fingerprinting of deep neural networks[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2019: 4729-4737.

[34] Yin Z, Yin H, Zhang X. Neural network fragile watermarking with no model performance degradation[C]//2022 IEEE International Conference on Image Processing (ICIP). IEEE, 2022: 3958-3962.

[35] Lin L, Wu H. Verifying integrity of deep ensemble models by lossless black-box watermarking with sensitive samples[C]//2022 10th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2022: 1-6.

[36] Botta M, Cavagnino D, Esposito R. NeuNAC: A novel fragile watermarking algorithm for integrity protection of neural networks[J]. Information Sciences, 2021, 576: 228-241.

[37] Abuadbba A, Rhodes N, Moore K, et al. Deepisign-g: Generic watermark to stamp hidden dnn parameters for self-contained tracking[J]. arXiv preprint arXiv:2407.01260, 2024.

[38] Yasui T, Tanaka T, Malik A, et al. Coded DNN watermark: robustness against pruning models using constant weight code[J]. Journal of Imaging, 2022, 8(6): 152.

[39] Yi X, Li Y, Zheng S, et al. Unified attacks to large language model watermarks: spoofing and scrubbing in unauthorized knowledge distillation[J]. Knowledge-Based Systems, 2025: 114295.

[40] Yang Z, Dang H, Chang E C. Effectiveness of distillation attack and countermeasure on neural network watermarking[J]. arXiv preprint

arXiv:1906.06046, 2019.

[41] Xiong Z, Wu G, Yu Q, et al. iSeal: Encrypted Fingerprinting for Reliable LLM Ownership Verification[J]. arXiv preprint arXiv:2511.08905, 2025.

[42] Jia J, Wang B, Gong N Z. Robust and verifiable information embedding attacks to deep neural networks via error-correcting codes[C]//Proceedings of the 2021 ACM Asia Conference on Computer and Communications Security. 2021: 2-13.

[43] Chattopadhyay N, Chattopadhyay A. Rowback: Robust watermarking for neural networks using backdoors[C]//2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, 2021: 1728-1735.

[44] Szyller S, Atli B G, Marchal S, et al. Dawn: Dynamic adversarial watermarking of neural networks[C]//Proceedings of the 29th ACM international conference on multimedia. 2021: 4417-4425.

[45] Tang R, Wigington C, Jin H, et al. Exposing Model Theft: A Robust and Transferable Watermark for Thwarting Model Extraction Attacks[C]//Proceedings of the 32nd ACM International Conference on Information and Knowledge Management. New York, NY, USA: Association for Computing Machinery, 2023: 4315-4319.

[46] Li H, Wenger E, Shan S, et al. Piracy resistant watermarks for deep neural networks[J]. arXiv preprint arXiv:1910.01226, 2019.

[47] Kapusta K, Thouvenot V, Bettan O, et al. A protocol for secure verification of watermarks embedded into machine learning models[C]//Proceedings of the 2021 ACM Workshop on Information Hiding and Multimedia Security. 2021: 171-176.

[48] Bansal A, Chiang P-Y, Curry M J, et al. Certified Neural Network Watermarks with Randomized Smoothing[C]//Proceedings of the 39th International Conference on Machine Learning. Baltimore, MD, USA: Proceedings of Machine Learning Research, 2022, 162: 1450-1465.

[49] 谭景轩, 钟楠, 郭钰生, 等. 深度神经网络模型水印研究进展[J]. 上海理工大学学报, 2024, 46(3): 225-242.

[50] Li Y, Zhang S, Wang W, et al. Backdoor attacks to deep learning models and countermeasures: A survey[J]. IEEE Open Journal of the Computer Society, 2023, 4: 134-146.

[51] Qiao Y, Liu D, Wang R, et al. Low-frequency black-box backdoor attack via evolutionary algorithm[C]//2025 IEEE/CVF Winter Conference on Applications of Computer Vision (WACV). IEEE, 2025: 7582-7592.

[52] Xie L, Liu X, Xie X, et al. IMUWatermark: A Blind and Robust Backdoor Watermark via Frequency-Domain Injection[C]//2025 IEEE 31th International Conference on Parallel and Distributed Systems (ICPADS). IEEE, 2025: 1-9.

[53] Liu Y, Wu H, Zhang X. Robust and imperceptible black-box DNN watermarking based on Fourier perturbation analysis and frequency sensitivity clustering[J]. IEEE Transactions on Dependable and Secure Computing, 2024, 21(6): 5766-5780.

[54] Lv P, Sun M, Wang H, et al. Rag-wm: An efficient black-box watermarking approach for retrieval-augmented generation of large language models[C]//Proceedings of the 2025 ACM SIGSAC Conference on Computer and Communications Security. 2025: 1709-1723.

[55] He J. A Novel Framework for Digital Image Watermarking Based on Neural Network[J]. Elektronika ir Elektrotechnika, 2024, 30(3): 83-93.

[56] Li Y, Maharana S K, Guo Y. Not just change the labels, learn the features: Watermarking deep neural networks with multi-view data[C]//European Conference on Computer Vision. Cham: Springer Nature Switzerland, 2024: 273-289.

[57] Fudenberg D, Tirole J. Game theory[M]. MIT Press, 1991.

[58] Tsai M J, Liu J, Wang C S. A game-theoretic framework for the security system of visible watermarking[J]. Expert Systems with Applications, 2011, 38(5): 5748-5754.

[59] Harsanyi J C. Games with incomplete information played by “Bayesian” players,

I–III Part I. The basic model[J]. Management Science, 1967, 14(3): 159-182.

[60] Nash Jr J F. Equilibrium Points in N-Person Games[J]. Proceedings of the National Academy of Sciences of the United States of America, 1950, 36(1): 48-49.

[61] Gupta S, Dube P, Verma A. Improving the affordability of robustness training for DNNs[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops. 2020: 780-781.

[62] Madaan D, Shin J, Hwang S J. Learning to generate noise for multi-attack robustness[C]//International Conference on Machine Learning. PMLR, 2021: 7279-7289.

[63] Pawlick J, Zhu Q. A Stackelberg game perspective on the conflict between machine learning and data obfuscation[C]//2016 IEEE International Workshop on Information Forensics and Security (WIFS). IEEE, 2016: 1-6.

[64] Li T, Li H, Pan Y, et al. Meta stackelberg game: Robust federated learning against adaptive and mixed poisoning attacks[J]. arXiv preprint arXiv:2410.17431, 2024.

[65] Dasgupta P, Collins J. A survey of game theoretic approaches for adversarial machine learning in cybersecurity tasks[J]. AI Magazine, 2019, 40(2): 31-43.

[66] Samsinger M, Merkle F, Schöttle P, et al. When Should You Defend Your Classifier? –A Game-Theoretical Analysis of Countermeasures Against Adversarial Examples[C]//International Conference on Decision and Game Theory for Security. Cham: Springer International Publishing, 2021: 158-177.

[67] Merkle F, Samsinger M, Schöttle P, et al. On the economics of adversarial machine learning[J]. IEEE Transactions on Information Forensics and Security, 2024, 19: 4670-4685.

[68] Graf J, Batchelor W, Harper S, et al. A practical application of game theory to optimize selection of hardware trojan detection strategies[J]. Journal of Hardware and Systems Security, 2020, 4(2): 98-119.

[69] Beltz B, Chen P Y, Doty J, et al. Guarding against malicious biased threats (gambit): Experimental design of cognitive sensors and triggers with behavioral

impact analysis[J]. arXiv preprint arXiv:2512.00098, 2025.

[70] Harsanyi J C. Games with incomplete information played by “Bayesian” players part II. Bayesian equilibrium points[J]. Management Science, 1968, 14(5): 320-334.

[71] Glicksberg I L. A further generalization of the Kakutani fixed point theorem, with application to Nash equilibrium points[J]. Proceedings of the American Mathematical Society, 1952, 3(1): 170-174.

[72] Gu W. Watermark removal scheme based on neural network model pruning[C]//Proceedings of the 2022 5th International Conference on Machine Learning and Natural Language Processing. 2022: 377-382.

[73] Aiken W, Kim H, Woo S, et al. Neural network laundering: Removing black-box backdoor watermarks from deep neural networks[J]. Computers & Security, 2021, 106: 102277.

[74] Guan J, Tu Z, He R, et al. Few-shot backdoor defense using shapley estimation[C]//Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. 2022: 13358-13367.

[75] Gu T, Dolan-Gavitt B, Garg S. Badnets: Identifying vulnerabilities in the machine learning model supply chain. arXiv 2017[J]. arXiv preprint arXiv:1708.06733, 2017.

[76] Li W, Wang G G, Gandomi A H. A Survey of Learning-Based Intelligent Optimization Algorithms: W. Li et al[J]. Archives of Computational Methods in Engineering, 2021, 28(5): 3781-3799.

[77] Atli B G, Xia Y, Marchal S, et al. WAFFLE: watermarking in federated learning[J]. arXiv preprint arXiv:2008.07298, 2020.

[78] Yang X, Xiang S, Peng C, et al. Federated learning incentive mechanism design via Shapley value and Pareto optimality[J]. Axioms, 2023, 12(7): 636.

[79] Li F Q, Wang S L, Liew A W C. Towards practical watermark for deep neural networks in federated learning[J]. arXiv preprint arXiv:2105.03167, 2021.

[80] Shapley L S. A Value for n-Person Games[C]//Kuhn H W, Tucker A W, eds. Contributions to the Theory of Games, Vol. 28. Princeton, NJ, USA: Princeton University Press, 1953: 307-317.

[81] Goodfellow I J, Pouget-Abadie J, Mirza M, et al. Generative adversarial nets[J]. Advances in Neural Information Processing Systems, 2014, 27.

[82] Sohl-Dickstein J, Weiss E, Maheswaranathan N, et al. Deep unsupervised learning using nonequilibrium thermodynamics[C]//International Conference on Machine Learning. PMLR, 2015: 2256-2265.

[83] Radford A, Kim J W, Hallacy C, et al. Learning transferable visual models from natural language supervision[C]//International Conference on Machine Learning. PMLR, 2021: 8748-8763.

[84] Alayrac J B, Donahue J, Luc P, et al. Flamingo: a visual language model for few-shot learning[J]. Advances in Neural Information Processing Systems, 2022, 35: 23716-23736.

[85] Heusel M, Ramsauer H, Unterthiner T, et al. Gans trained by a two time-scale update rule converge to a local nash equilibrium[J]. Advances in Neural Information Processing Systems, 2017, 30.

[86] Hessel J, Holtzman A, Forbes M, et al. Clipscore: A reference-free evaluation metric for image captioning[C]//Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing. 2021: 7514-7528.

[87] Lansari M, Bellafqira R, Kapusta K, et al. When federated learning meets watermarking: A comprehensive overview of techniques for intellectual property protection[J]. Machine Learning and Knowledge Extraction, 2023, 5(4): 1382-1406.

攻读硕士学位期间取得的研究成果

一、论文

[1] **Huang C**, Wu H. A game between the defender and the attacker for trigger-based black-box model watermarking[C]//2025 13th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2025: 1-6.**(EI: 20250029690)**

[2] **Huang C**, Zhao G, Wu H, et al. An information asymmetry game for trigger-based DNN model watermarking[C]//2026 14th International Symposium on Digital Forensics and Security (ISDFS). IEEE, 2026: 1-6.**(EI: 20250450365)**

[3] Zhang Y, **Huang C**, Liu S, et al. Screen-shooting resistant robust document watermarking in the Discrete Fourier Transform domain[J]. International Journal of Network Management, 2025, 35(1): e2278.**(WOS:001237369300001, EI: 20242316203470)**

致　谢

在论文完成之际，我想由衷感谢在硕士研究生阶段关心与帮助过我的人。

首先，感谢我的导师吴汉舟老师，在硕士阶段的三年学习和生活中，吴老师不仅以身作则督促着我刻苦钻研、积极进取，还在我的论文选题和论文撰写的过程中对我进行了悉心的指导和耐心的帮助。

其次，我还要感谢实验室的同门以及师兄师姐和师弟师妹们，感谢他们在科研道路上和生活方面给我提供的帮助，感谢他们在实验室一起经历科研的过程中为我提供的答疑帮助。因为有了他们，我的研究生三年生活变得丰富多彩。在这里衷心祝愿他们未来一切顺利。与大家在实验室一起奋斗的时光对于我的人生来说是一笔宝贵的财富。

然后，我还要感谢我的家人和朋友们。在我的研究生生涯中，他们始终是我最坚实的后盾，有了他们的支持我才有底气去面对和挑战遇到的所有难题。无论我遇到什么样的挫折和困难，他们始终给予我最大的鼓励与帮助，让我能够全身心的投入到研究生学术研究中去。

最后，感谢百忙之中抽出时间评阅论文的专家老师们，您们的建议和意见对我的学术研究和人生规划都有很大的帮助。在未来的道路上，我会一如既往地严谨、努力，争取取得更多的成就。

黄超越

上海大学

2026 年 5 月 10 日