目录

1 CS 285 深度学习强化学习—第 15 讲详细讲义

原文来源：lec-15.pdf **生成日期**：2026-05-09 **总页数**：22 页 | **总章节**：3 章 **语言**：简体中文 **课程名称**：Model-Based RL（基于模型的强化学习）**讲师**：Sergey Levine, UC Berkeley **说明**：本讲义基于课程幻灯片深度扩写而成，每页幻灯片均扩展为详细讲解，补充了背景知识、公式推导、实例分析和关键要点总结。

1.1 目录

1. 第 1 章：我们能学习一个”模拟器”吗
2. 第 2 章：分布偏移与不确定性
3. 第 3 章：不确定性感知神经网络

1.2 第 1 章：我们能学习一个”模拟器”吗

涵盖范围：Slide 2-8 | **核心主题**：探讨”学习环境模拟器并用它来进行规划/RL”这一想法的可行性、局限性与挑战

1.2.1 1.1 学习模拟器的愿景（Slide 2-4）

概念详解

Slide 2-3 用一个大胆的问题开启了基于模型的强化学习（Model-Based RL）的讨论：“我们能学习一个模拟器吗?”（Can we learn a “simulator”?）Slide 3 展示了 Veo 2 和 Sora 等视频生成模型——这些模型能够从文字描述生成逼真的物理世界视频，暗示了大神经网络学习复杂环境动态的潜力。

“学习模拟器”的愿景是诱人的：如果我们可以从数据中学习一个足够准确的环境模拟器（world model），那么我们就可以在这个学习到的模拟器中进行”虚拟规划”——无需在真实环境中执行代价高昂的试错。这种想法在直觉上非常自然：人类在做复杂决策之前会在脑海中进行”心理模拟”（mental simulation），为什么不让 AI 也这样做?

Slide 4 展示了一个概念性的”学习模拟器”算法流程：收集数据 → 学习动力学模型 $\hat{p}(s_{t+1}|s_t, a_t)$ → 在学到的模型中运行 model-free RL 或规划算法 → 将学到的策略部署到真实环境。这个流程看似简单直观，但 Slide 4 以一个尖锐的问题收尾：“这会有效吗？为什么可能会失败?”（Will this work well? Why might it fail?）

深度剖析

“学习模拟器”这一愿景面临的核心矛盾是：**要学习一个好的模拟器，你需要覆盖所有可能状态-动作区域的数据；但要获取这些数据，你需要在真实环境中探索——而这恰恰是学习模拟器想要避免的代价**。

这构成了一种”先有鸡还是先有蛋”的困境：- 如果我们只有少量数据 → 学到的模拟器只在这些数据覆盖的区域内是准确的 - 如果我们在模拟器中规划/优化 → 策略自然会引导智能体去模拟器尚未见过的区域（因为在那里模拟器的预测可能是过度乐观的——这正是我们即将讨论的分布偏移问题的根源）- 如果我们在这些新区域真实执行 → 可能带来灾难性的失败

此外，还有一个**计算效率**的问题：即使模拟器完全准确，在一个高维观测空间（如图像像素）上进行规划或 RL 也是极其困难的。传统的规划算法（如 MCTS、iLQR）在像素空间中几乎不可用——它们需要在抽象的状态空间中进行推理。

关键要点

- 学习模拟器的愿景：从数据中学环境动态 → 在模拟器中规划 → 降低真实交互成本
- Veo 2、Sora 等视频生成模型展示了学习复杂动态的可能性
- 核心矛盾：准确学习需要广泛数据，但数据获取正是模拟器想要规避的代价
- 像素空间中的规划是另一个独立的技术难题

  过渡衔接：让我们更精确地分析”学习模拟器”方法可能失败的原因——这将引出 model-based RL 中最根本的挑战。

1.2.2 1.2 为什么学习模拟器会失败（Slide 5-8）

概念详解

Slide 5-7 系统地剖析了学习模拟器方法的潜在失败模式。

失败模式 1：策略在模拟器中学到的行为可能严重依赖模拟器的误差。这是分布偏移（distributional shift）的核心表现。如果学到的动力学模型在某个区域有系统性的误差（例如，低估了某个动作的风险），策略优化器会自然地”利用”

这个误差——引导智能体去做在模拟器中看起来很好但在现实中会失败的动作。Slide 5 指出”this policy matters a lot”——模拟器中的最优策略可能是真实环境中的灾难策略。

失败模式 2：模型设计是领域相关的。 Slide 6 用了一个生动的例子：视频生成模型可以生成”一个类人机器人站在桌子旁，桌上放着红绿蓝三色方块，正在执行堆叠任务，红色在底部，蓝色在顶部”的逼真视频——但这并不意味着模型真正理解物理约束。模型可能只是在视觉上”看起来合理”，而忽略了关键物理细节（如接触力、摩擦力、稳定性）。

失败模式 3：大神经模型在计算上更昂贵。与传统的物理引擎（如 MuJoCo、Isaac Gym）相比，基于深度学习的模拟器推理成本高出数个数量级。在实践中，使用学到的模拟器进行数千次模拟（这在规划中是必要的）可能不可行。

深度剖析

Slide 8 将这些挑战组织为一个三层框架：

1. **统计/算法问题（Statistics/algorithms problem）**：“课上讨论起来很有趣，因为我们对这层有很好的理解”——这是学术界最熟悉的层面。如何处理分布偏移？如何在模型不确定的区域做出安全的决策？如何平衡探索和利用？

2. **深度学习/模型问题（Deep learning/models problem）**：“工业界讨论起来很有趣，因为我们可以让 GPU 嗡嗡作响”——这是以算力驱动的进步。如何设计更好的世界模型架构？如何扩展到更大规模的数据和模型？

3. **控制/RL 问题（Controls/RL problem）**：“讨论起来通常没那么有趣，因为我们已经知道该怎么做”——这是最成熟但也最容易在实践中被忽视的层面。如何在学到的（不完美的）模型中进行有效规划？如何选择规划算法？

这个三层框架的一个重要洞察是：**第一层和第三层常常被”更多的数据和更强的模型”所掩盖，但它们不会因此消失。**再好的世界模型，如果策略优化器利用其微小误差来”作弊”，最终在真实环境中的表现仍然会很差。

关键要点

- 策略会利用模拟器的误差——分布偏移导致”模拟器中好，现实中差”
- 模型可能”看似合理”但忽略了关键物理细节
- 深度学习模拟器的推理成本远高于传统物理引擎
- 三层框架：统计算法层 + 深度学习模型层 + 控制规划层，缺一不可

过渡衔接：分布偏移是 model-based RL 最根本的挑战。下一章将详细分析分布偏移的本质和潜在的解决策略。

1.3 第 2 章：分布偏移与不确定性

涵盖范围：Slide 9-15 | **核心主题**：深入分析 model-based RL 中分布偏移的本质、为什么简单的解决方法不够、以及不确定性估计如何提供帮助

1.3.1 2.1 分布偏移的本质（Slide 9-12）

概念详解

Slide 9-12 深入探讨了分布偏移（Distributional Shift）在 model-based RL 中的表现。分布偏移指的是：模型是在某些状态-动作分布上训练的，但策略可能会引导智能体进入模型从未见过的状态区域——在那里模型的预测是不可靠的。

Slide 10 展示了一个经典的教学场景：智能体学到的模型在已探索区域（训练分布内）表现良好，但一旦策略试图"走出去"探索新区域，模型就开始产生错误的预测。更重要的是，模型在未见区域产生的**过度自信的错误预测**往往会制造"海市蜃楼"——模型可能错误地预测高奖励，引诱策略离开安全区域。

Slide 11 尖锐地提出了一个两难困境："但我们应该走多远?"（But how far do we go?）："如果我们想学得快，就应该尽可能走远……"但走得越远，模型越不可靠。这是 model-based RL 中**探索-利用**权衡在模型不确定性维度的体现。

Slide 12 用一句直白的话总结了问题的严重性："这问题相当糟糕!"（The problem is pretty bad!）—— "很容易想去这里……"（very tempting to go here…）——指的正是模型在未见区域制造的虚假高奖励预期。

深度剖析

分布偏移为什么在 model-based RL 中比在 model-free RL 中更加危险? 在 model-free RL 中，Q 函数或策略的过估计（overestimation）虽然也存在，但通常可以通过技术手段（如 Double Q-learning、Clipped Double Q-learning）来缓解。但在 model-based RL 中，动力学模型的误差会被**复合放大**：

1. 策略规划器利用模型来"模拟"多步轨迹
2. 每一步的模拟都引入了微小的模型误差
3. 在多步滚动（rollout）中，这些误差累积——第一步的小误差导致策略做出了一个略微不同的动作选择，这个动作又落在模型更不熟悉的区域，产生更大的误差……
4. 最终，规划出的轨迹与真实环境中的最优轨迹可能截然不同

这种误差的复合放大效应是 model-based RL 区别于其他范式的最核心的挑战。它解释了为什么即使模型在"一步预测"上的准确率很高（如 $>95\%$），多步规划仍然可能失败。

关键要点

- 分布偏移 = 策略进入模型未见过的区域 $\rightarrow$ 模型预测不可靠

- “海市蜃楼”效应：模型在未知区域错误地预测高奖励，引诱策略离开安全区
- 探索-利用在 model-based RL 中的新维度：走多远才能平衡学习速度和模型可靠性?
- 模型误差在规划中复合放大——一步误差小不等于多步规划可靠

过渡衔接：面对分布偏移，Slide 13 给出了解决策略的两条主线——限制策略的变化程度，以及让模型感知自身的不确定性。

1.3.2 2.2 解决分布偏移的策略（Slide 13-14）

概念详解

Slide 13 概述了应对分布偏移的两大类策略：

策略 1：不要过多改变策略（Don’ t change the policy too much）。这是”信任区域”（trust region）方法的思路。通过在策略更新中施加约束（如 KL 散度约束），确保新策略不会偏离旧策略收集数据的分布太远。这与 TRPO/PPO 的思想一脉相承——在 model-based RL 中，信任区域不仅防止策略崩溃，还确保策略不进入模型不可靠的区域。

Slide 13 列出了更具体的方法：- **使用概率性的、不确定性感知的模型**：模型不仅给出预测，还给出”我对这个预测有多确定”- **在模型不确定的区域使用惩罚（悲观主义）**：给不确定区域的预测施加”悲观修正”——宁可保守地低估奖励，也不乐观地高估

策略 2：坚持模型”确信”的区域（Stick to places where the model is confident）。这要求模型能够量化自身的不确定性——知道哪些区域是熟悉的、哪些是陌生的。

深度剖析

Slide 14 用数学直观说明了为什么不确定性估计有帮助。考虑两个预测：预测 A 的期望奖励为 10，但方差很大（模型高度不确定）；预测 B 的期望奖励也为 10，但方差很小（模型很确定）。虽然期望值相同，但**悲观估计**（pessimistic estimate，如期望值减去若干标准差）会给出完全不同的结果：

$$\text{悲观值}_A = 10 - \alpha \cdot \sigma_{\text{high}} \ll \text{悲观值}_B = 10 - \alpha \cdot \sigma_{\text{low}}$$

这自然引导策略选择预测 B（在模型确定的区域行动），避开预测 A（在模型不确定的区域冒险）。这种”不确定即惩罚”的机制优雅地解决了分布偏移问题——它不需要显式地约束策略的探索范围，而是通过奖励信号自动引导策略远离不确定性。

这种方式被称为**悲观主义**（Pessimism）或**保守主义**（Conservatism）原则：在面对不确定性时，宁可保守行事。这在理论上与离线 RL（offline RL）中的保守 Q 学习（Conservative Q-Learning, CQL）以及稳健控制理论中

的最坏情况优化（worst-case optimization）共享相同的哲学基础。

关键要点

- 两类策略：约束策略变化（信任区域）vs 感知并回避不确定性
- 概率模型 + 悲观惩罚：不确定区域的预测自动被降权
- 悲观值 = 期望值 $-\alpha \cdot$ 不确定性：不确定性越高，修正幅度越大
- 悲观主义原则在理论上与稳健控制和离线 RL 共享哲学基础

过渡衔接：悲观主义依赖于对不确定性的准确估计。但 Slide 15 提出了一个重要的修正——有时候我们不仅需要”避开不确定区域”，还需要主动探索它们。

1.3.3 2.3 探索与乐观主义的角色（Slide 15）

概念详解

Slide 15 指出了悲观主义原则的一个重要局限：“需要探索才能变得更好”（Need to explore to get better）。纯粹的悲观主义会让智能体永远停留在模型确定的区域——但那些区域可能只有次优的奖励。为了发现更好的策略，智能体有时必须进入不确定的区域。

Slide 15 用三个陈述精确刻画了不同估计方式在探索-利用中的角色：

- **期望值不等于悲观值（Expected value is not the same as pessimistic value）**：悲观值系统性地低估了在不确定区域的真实收益
- **期望值不等于乐观值（Expected value is not the same as optimistic value）**：乐观值会引诱智能体过度冒险
- ……**但期望值通常是一个好的起点（but expected value is often a good start）**：在实践中，不应该完全依赖悲观主义或乐观主义，而是在期望值的基础上进行适度的修正

深度剖析

探索与保守之间的张力在 model-based RL 中以一种新的形式呈现。在 model-free RL 中，探索通常通过 ϵ-greedy、熵正则化或内在奖励（intrinsic reward）来驱动。在 model-based RL 中，探索有了更丰富的理论基础——它可以被形式化为**减少模型不确定性的信息收集行为**。

具体来说，如果模型在一个区域高度不确定，那么：- **悲观视角**：避免这个区域（因为可能有害）- **乐观视角**：进入这个区域（因为可能有益，而且探索后模型变得更确定，长期收益更高）

最优策略应该在两者之间取得平衡——这被称为 **“面对不确定性的乐观主义”**（Optimism in the Face of Uncertainty, OFU）。OFU 原则在 bandit 问题和 tabular RL 中有坚实的理论保障（如 UCB 算法），但在深度

model-based RL 中的实现仍然是一个活跃的研究领域。

Slide 15 的告诫" 有一些注意事项…… "暗示了纯乐观主义的风险：在高维连续空间中，不确定区域可能是无穷的，不加选择地探索所有不确定区域在计算上不可行，且可能带来灾难性的后果。

关键要点

- 纯悲观主义导致智能体永不探索——需要在安全和探索间平衡
- 悲观值　期望值　乐观值——每种估计有不同的探索-利用含义
- 面对不确定性的乐观主义（OFU）在理论上优雅但在实践中需谨慎
- 期望值作为起点，辅以适度的悲观修正，是当前实践中的常见策略

> **过渡衔接**：无论是悲观主义还是乐观主义，前提都是模型能够量化自身的不确定性。这引出了第三部分的核心问题：如何构建不确定性感知的神经网络?

1.4 第 3 章：不确定性感知神经网络

> **涵盖范围**：Slide 16-22 | **核心主题**：构建能够量化预测不确定性的神经网络模型——从偶然不确定性与认知不确定性的区分，到贝叶斯神经网络和 Bootstrap Ensembles 的实用方法

1.4.1 3.1 不确定性的两种类型（Slide 16-17）

概念详解

Slide 16-17 区分了机器学习中两种根本不同的不确定性类型：

偶然不确定性（Aleatoric Uncertainty）：也称为统计不确定性（statistical uncertainty）。它源于数据本身的固有随机性——即使我们有无限多的数据和一个完美的模型，这种不确定性也不会消失。例如，掷一枚公平硬币的结果就具有偶然不确定性——没有人能确定地预测下一次是正面还是反面。Slide 17 问" 这里的方差是什么? "(what is the variance here?）——指的正是输出分布本身的方差。

认知不确定性（Epistemic Uncertainty）：也称为模型不确定性（model uncertainty）。它源于我们对模型参数的知识不足——"模型对数据很确定，但我们对模型不确定"（the model is certain about the data, but we are not certain about the model）。随着我们收集更多数据，认知不确定性可以减少；极限情况下（无限数据），认知不确定性可以趋近于零。

Slide 17 给出了一个精辟的区分："模型对数据很确定，但我们对模型不确定"——这句话区分了" 模型知道自己在做什么" 和" 我们（作为模型的使用者）知道模型在做什么"。

深度剖析

理解这两种不确定性的区分对于 model-based RL 至关重要，因为它们需要被**区别对待**：

偶然不确定性（数据固有的噪声）：在 model-based RL 中，这部分不确定性是环境动态本身的一部分——即使我们完全知道转移概率，状态转移仍然是随机的。这对应于环境动态中的不可约噪声。当我们在环境中执行相同动作多次但得到不同结果时，那就是偶然不确定性在起作用。

认知不确定性（模型知识的不足）：这是我们最关心的部分——它告诉我们模型在哪些区域缺乏训练数据。在那些区域，模型的预测不应该被信任。随着收集更多数据，认知不确定性降低，模型预测变得可靠。

Slide 17 指出仅使用输出熵（output entropy）来估计不确定性”是不够的”（why is this not enough?）。因为输出熵混合了两种不确定性——它不能区分”因为数据本身是随机的而导致高熵”和”因为模型不确定而导致高熵”。在 model-based RL 中，我们需要的是**只对认知不确定性做悲观修正**——不应该因为环境的固有随机性（偶然不确定性）而惩罚探索。

关键要点

- 偶然不确定性 = 数据固有的随机性，不可约
- 认知不确定性 = 我们对模型参数的知识不足，可随数据增多而减少
- 输出熵混合了两种不确定性——不能替代真正的认知不确定性估计
- 在 model-based RL 中需要对认知不确定性（而非偶然不确定性）进行悲观修正

过渡衔接：区分了两种不确定性之后，下一个问题是：如何实际估计认知不确定性？Slide 18-20 给出了几种可行的方法。

1.4.2 3.2 估计模型不确定性（Slide 18-20）

概念详解

Slide 18 提出了估计模型不确定性的核心思想：**模型的”不认同”程度**（how much the models disagree）。如果我们有多个独立的模型，它们对同一输入给出不同的预测，那么这些预测之间的分歧程度就反映了认知不确定性。在数据充足的区域，所有模型应该给出相似的预测（低分歧）；在数据稀疏的区域，不同模型可能给出截然不同的预测（高分歧）。

Slide 19 简要介绍了贝叶斯神经网络（Bayesian Neural Networks, BNN）的基本思想。在 BNN 中，模型的权重不是确定的点估计，而是**概率分布**——每个权重 θ_i 有一个均值和方差。预测时，我们从权重分布中采样多个模型，这些样本之间的分歧量化了认知不确定性。关键的数学概念：

$$\text{权重} \sim p(\theta|\mathcal{D}) \quad \text{而非} \quad \theta = \theta_{\text{MAP}}$$

Slide 19 引用了 Blundell et al.（Weight Uncertainty in Neural Networks）和 Gal et al.（Concrete Dropout）作为 BNN 方法的深入参考文献。

深度剖析

贝叶斯神经网络在理论上是优雅的，但在实践中面临严重的可扩展性问题：- 完整的贝叶斯推断需要对权重的高维后验分布进行积分——这在大型神经网络中是不可行的 - 近似方法（如变分推断、MC Dropout）提供了可扩展的替代方案，但各有其局限

Slide 20 介绍了一个在实践中更常用的替代方案：**Bootstrap Ensembles**（自助集成）。核心思想是：1. 从原始数据集中通过自助采样（bootstrap sampling，有放回重采样）生成 K 个”独立”的数据集 2. 在每个数据集上训练一个独立的模型 3. 这些模型之间的分歧量化了认知不确定性

在深度学习中，Slide 21 指出了一些实用的简化：- **(自）重采样通常是不必要的**：因为 SGD 的随机性和随机初始化已经使模型之间足够”独立”- **模型数量通常很少（$<$ 10)**：这是一个粗糙的近似，但在实践中效果不错 - **每个模型独立地随机初始化和 SGD 训练**：这本身就引入了足够的多样性

关键要点

- 模型分歧（disagreement）是衡量认知不确定性的实际代理指标
- 贝叶斯神经网络理论上优雅但实践中难以扩展
- Bootstrap Ensembles：训练 K 个独立模型，用它们的预测分歧度量不确定性
- 深度学习中重采样不是必须的——SGD 随机性 + 随机初始化已经提供足够的模型多样性
- 通常 K $<$ 10 就足够在实践中发挥作用

过渡衔接：有了不确定性感知的模型和悲观主义的原则，下一次课将把这些组件整合为完整的 model-based RL 算法。Slide 22 为本讲画上了句号，也为下一讲埋下了伏笔。

1.4.3 3.3 总结与展望（Slide 22）

概念详解

Slide 22 作为第 15 讲的收尾,预告了下一次课的内容:“使用不确定性感知的模型来处理分布偏移”(use uncertainty-aware model to deal with distributional shift)。这暗示着第 15 讲和第 16 讲之间的自然衔接：本讲建立了问题（分布偏移）和工具（不确定性估计），下一讲将展示如何使用这些工具来构建实用的 model-based RL 算法。

Slide 22 还提示：“可以运行我们学过的某一种算法，但有更好的选择”（could run one of the algorithms we learned about, but there are better options here）。这指的是：理论上，我们可以将 model-based RL 简单地实现为”学模型 → 在模型中使用 model-free RL”，但这样做效率不高——有专门为 model-based 场景设计的更优算法（如基于模型的策略优化 MB-MPO、概率集成与轨迹采样 PETS 等，将在后续讲座中介绍）。

深度剖析

回顾整堂课的三个部分，我们可以看到一个清晰的方法论演进：

1. **识别问题**（第 1 章）：学习模拟器的愿景是诱人的，但分布偏移使其在实践中危机四伏
2. **建立框架**（第 2 章）：分布偏移可以通过信任区域约束和不确定性感知来应对——悲观主义在不确定区域提供”安全网”
3. **提供工具**（第 3 章）：不确定性感知可以通过 Bootstrap Ensembles 等实用方法实现——虽然贝叶斯方法的全貌是深奥的，但我们有简单有效的近似

这一框架奠定了后续 model-based RL 课程的基石。特别是，**将动力学模型的不确定性估计与策略优化相结合**这一思想，将在后续的讲座中反复出现——从基于模型的价值扩展（MVE）到 Dreamer 系列算法。

关键要点

- 下一讲将整合不确定性感知模型与策略优化
- 简单的”学模型 + 使用 model-free RL”虽然可行但效率不高
- 专门的 model-based RL 算法设计利用了模型的独特结构
- 不确定性是 model-based RL 的核心概念——它贯穿于探索、规划和策略优化全过程

本讲总结：第 15 讲开启了 CS 285 课程的 model-based RL 板块。从”学习模拟器”的美好愿景出发，课堂系统地揭示了这一范式面临的根本性挑战——分布偏移，以及应对这一挑战的核心思路——不确定性感知。偶然不确定性与认知不确定性的区分、贝叶斯神经网络与 Bootstrap Ensembles 的权衡、悲观主义原则与探索需求之间的张力——这些构成了 model-based RL 的方法论语料库。本讲提出的问题多于答案，但这些问题正是驱动后续讲座深入展开的动力。