(paper)量化非平稳性对基于强化学习的交通信号控制的影响

摘要：

> 在强化学习（RL）中，处理非平稳性是一个具有挑战性的问题。 然而，某些领域（例如交通优化）本质上是非平稳的。 造成这种情况的原因和影响是多方面的。 特别是，在处理交通信号控制时，解决非平稳性是关键，因为交通状况随着时间的推移而变化，并且是网络其他部分所采取的交通控制决策的函数。 在本文中，我们分析了不同的非平稳性来源对交通信号网络的影响，其中每个交通信号控制机都被建模为学习智能体。 更准确地说，我们研究改变智能体学习环境的影响（例如，其经历的交通流变化），以及降低智能体对真实环境状态的可观察性的影响。 部分可观察性可能会导致交通信号智能体将不同的状态（其中是不同的最佳操作）视为相同。 反过来，这可能会导致性能不佳。 我们发现，缺乏合适的传感器来提供对真实状态的代表性观察似乎比底层流量模式的变化对性能的影响更大。

1，介绍

> 控制交通信号是处理使用现有城市网络基础设施的车辆数量不断增加的一种方法。强化学习(RL)通过允许去中心化(交通信号-建模为代理-可以独立学习在每个当前状态下采取的最佳行动)以及对交通流量变化的动态适应，增加了这种努力。值得注意的是，这可以通过RL技术以无模型的方式完成(没有预先的领域信息)。强化学习是基于一个代理来计算将状态映射到动作的策略，而不需要显式的环境模型。这在交通领域很重要，因为这样的模型可能非常复杂，因为它涉及到

> 建模交通状态转换，不仅由多个代理的动作决定，而且还由环境固有的变化决定，例如车辆流量的时间依赖性变化。

> 在交通控制问题中应用强化学习(RL)的主要困难之一是环境可能以不可预测的方式变化。代理可能必须在不同的上下文中运行——我们在这里将其定义为影响代理的真正底层流量模式;重要的是，智能体不知道其环境的真实背景，例如，因为它们不具有交通网络的完全可观察性。导致不同环境的部分可观察变量的例子包括一天中不同时段的不同交通模式、交通事故、道路维护、天气和其他危险。我们把环境动力学的变化称为非平稳性。

> 然后，我们分析了不同的非平稳性来源——当将强化学习应用于交通信号控制时——并量化了每种来源在学习过程中的影响。更准确地说，我们研究了以下因素对学习绩效的影响:(1)不同车辆流量引起的交通模式的显式变化;(2)由于交叉口传感器的读数不精确或不可用，导致状态可观测性降低。

> 后一个问题可能会导致不同的状态(其中不同的行为是最优的)被交通信号代理视为相同的。这不仅会导致次优性能，而且在环境上下文发生变化时可能会导致性能急剧下降。

> 我们评估了在非平稳多智能体场景中部署强化学习的性能，其中每个交通信号都使用q -学习(一种无模型强化学习算法)来学习有效的控制策略。

> 使用开源微观交通仿真系统SUMO (Simulation of Urban MObility)对交通环境进行模拟，并对具有16个交通信号智能体的4 × 4网格交通网络的动态建模，其中每个智能体只能访问其受控交叉路口的本地观测值。

> 我们的实验证明，上述非平稳性的原因会对学习智能体的性能产生负面影响。我们还证明，缺乏合适的传感器来提供真实底层交通状态的代表性观察，似乎比底层交通模式的变化对学习性能的影响更大。

2.2 RL的非平稳性

在强化学习中，处理非平稳性是一个具有挑战性的问题。非平稳性的主要原因包括状态转移函数T (s, a, s′)或奖励函数R(s, a, s′)的变化、真实环境状态的部分可观察性以及其他智能体所采取行动的不可观察性。

在MDP中，假设状态转移函数T不变。然而，这在许多现实问题中是不现实的。在非平稳环境中，状态转移函数T和/或奖励函数R可以在任意时间步长变化。例如，在交通领域中，给定状态下的动作可能会根据当前环境产生不同的结果，网络环境会随着智能体的动作而变化。如果智能体没有显式地处理环境的变化，它们可能不得不重新调整策略。因此，他们可能会经历一个持续的遗忘和重新学习控制策略的过程。虽然这种重新适应是可能的，但它可能会导致智能体在较长时间内以次优方式运行。

2.3 部分可观测性

> 交通控制问题可以建模为Dec - POMDP——一种特殊类型的分散多智能体MDP模型——智能体对环境的真实状态只有部分可观察性。

在交通信号控制中，由于缺乏合适的传感器来提供交通交叉口的代表性观测，可能会出现部分可观测性。此外，即使有多个传感器可用，由于测量不准确(低分辨率)，也可能出现部分可观测性。

3，方法

本文的主要目标是：交通智能体学习如何在各种形式的非平稳性下改善交通流，在这个场景中，研究可能影响性能的非平稳性的不同原因。

> 为了研究这一问题，我们引入了一个时变动态下的城市交通建模框架。特别是，我们首先介绍了一个基于mdp的基线城市交通模型。这是通过形式化MDP的相关元素(遵循类似的现有工作)来实现的:它的状态空间、动作集和奖励函数。

然后，我们将展示如何扩展此基线模型，以允许对其转态转移函数进行动态更改，从而对不同环境的存在进行编码。

在这里，环境对应于不同的流量模式，这些模式可能会随着时间的推移而变化，而这些变化的原因可能无法被智能体直接观察到。

我们还讨论了关于交通系统状态定义的可能方式的不同设计决策;其中许多与文献中典型的建模选择一致。

讨论状态的不同可能定义是相关的，因为这些定义通常以直接包含传感器信息的方式指定。

然而，考虑到传感器信息的数量和质量，不同的状态定义会产生不同的非平稳性，这取决于传感器分辨率和环境和/或其他智能体的部分可观测性。

此外，在接下来的内容中，我们描述了每个交通信号智能体使用的多智能体训练方案(在第3.4节中)，以便在非平稳性设置下优化其策略。

我们还描述了交通模式——我们的代理可能需要操作的环境——是如何被数学建模的(第3.5节)。

我们在第4节中讨论了用于分析和量化交通问题中非平稳性影响的方法。

最后，我们在此强调，本文中提出的方法和分析——旨在评估不同非平稳来源的影响——是我们工作的主要贡献。

大多数现有的工作(例如，第5节中讨论的那些工作)都没有解决或直接详细调查交通流量变化作为RL非平稳性来源的影响。

3.1 状态构成

在我们处理的问题或场景中，状态空间的定义强烈地影响着智能体的行为和性能。每个交通信号智能体控制一个十字路口，在每个时间步t，它观察一个向量st，该向量st部分表示被控制十字路口的真实状态。

在我们的问题中，一个状态可以定义为向量s∈R(|P|)，其中P是所有信号相位的集合，ρ∈P表示当前相位（绿灯），δ∈[0,maxGreenTime]是当前相位经过的时间，Density(i)∈[0,1]定义为车辆密度，queue(i)∈[0,1]定义为排队长度（密度）。

$$s = [\rho, \delta, density_1, queue_1, ..., density_{|P|}, queue_{|P|}]$$

请注意，由于许多物理传感器需要付费和部署，因此存在成本问题，因此这种状态定义在现实环境中可能不可行。出于这个原因，我们引入了另一种状态定义，它缩小了观察的范围。更准确地说，这种替代状态定义从上式中删除了密度属性，导致下式中的部分可观察状态向量s∈R(|P |)。缺乏这些状态属性类似于缺乏现实生活中的交通传感器，这些传感器能够检测沿着给定街道延伸的驶入车辆(即沿着该街道的车辆密度)。

$$s = [\rho, \delta, queue_1, ..., queue_{|P|}]$$

还要注意，上述定义的结果是连续状态。然而，传统上，Q-learning处理离散状态空间。因此，状态计算后需要离散化。密度和队列属性都被离散到十个平均分布的层/箱中。我们指出，低水平的离散化也是部分可观测性的一种形式，因为它可能导致不同的状态被认为是相同的状态。此外，在本文中，我们假设——正如文献中通常做的那样——一个模拟时间步长对应于现实生活中5秒的交通动态。这有助于编码交通信号通常不会每秒钟改变动作的事实;这个建模决策意味着动作(特别是对交通灯当前相位的更改)每隔5秒执行一次。

3.2 动作

在MDP中，在每个时间步t，每个智能体选择一个动作a∈A，在我们的设置中，动作的数量等于相位的数量，其中一个相位打开绿灯信号放行一个方向的交通流，因此，|A| = |P |。在交通网络为网格的情况下，我们考虑两种行为:智能体可以保持当前相位的绿灯时间，也可以允许绿灯时间进入另一个相位;我们分别称这两个动作为“保持”和“改变”。在动作选择中有两个限制:只有当δ≥10s (minGreenT time)时，智能体才能采取动作改变;只有当δ < 50s (maxGreenT time)时，智能体才能保持动作。

另外，改变动作会施加一个固定持续时间为2秒的黄色相位。这些限制的存在是为了模拟现实生活中的一个事实，即交通控制器需要在最短的时间内做出决定，以允许停止的汽车加速并驶向预定目的地。

3.3 回报函数

在我们的模型中，分配给交通信号智能体的奖励被定义为连续动作之间累积车辆等待时间的变化。在执行一个动作后，智能体得到的奖励r(t)∈R，如下式所示:

$$r_t = W_t - W_{t+1}$$

式中，W(t)和W(t+1)表示执行动作a(t)前后在路口的累计等待时间;下式:

$$W_t = \sum_{v \in V_t} w_{v,t}$$

式中Vt为时间步长为t时到达交叉口的道路上的车辆集合，w(v,t)为车辆v进入交叉口入口的道路至时间步长为t时的总等待时间。如果车速低于0.1 m/s，则认为车辆处于等待状态。注意，根据这个定义，累积等待时间减少越多，奖励就越大。因此，通过最大化奖励，交通智能体减少了车辆在十字路口的等待时间，从而改善了交叉口的交通状态。

3.4 多智能体独立q学习

我们通过在多智能体独立训练方案中使用Q学习来解决我们场景中的非平稳性问题，其中每个交通信号机都是一个QL智能体，具有自己的Q表、局部观察、动作和奖励。这种方法允许每个智能体学习一个单独的策略，适用于它所做的局部观察;智能体之间的策略可能不同，因为每个智能体仅使用自己的经验元组更新其Q表。除了允许智能体之间的不同行为外，这种方法还避免了集中式训练方案可能引入的维度问题。然而，独立训练方案有一个主要缺点:当智能体学习和调整它们的策略时，它们的策略的变化会导致环境状态变化，从而导致非平稳。这意味着单智能体算法的原始收敛特性不再成立，因为一个智能体的最佳策略会随着其他智能体策略的变化而变化。

3.5 情境

为了给环境中导致不平稳性的因素建模，我们使用了交通情境的概念，类似于参考文献中的方法[12]。我们将情境定义为由交通网络的(OD)对上的不同车流分布组成的交通模式。OD对的原始节点表示车辆在仿真中的插入位置。目的地节点是车辆结束其行程的节点，因此在车辆到达时将其从仿真中移除。然后，通过将每秒插入到原始节点的若干车辆与每个OD对相关联来定义交通情境。

在仿真过程中改变情境会导致传感器测量值随时间的变化而变化。例如，交通事故和高峰时段等事件会导致某一方向的车流量增加，从而使该方向车道上的队列增加得更快。在通常情况下，当智能体不能访问环境的状态的所有信息时，这会直接影响MDP的状态转移T和奖励函数R。因此，当状态转移概率和奖励发生变化时，状态-行为对的Q值也会发生变化。因此，交通信号智能体很可能需要经历一个重新适应阶段，以正确地更新它们的策略，从而导致性能的灾难性下降。

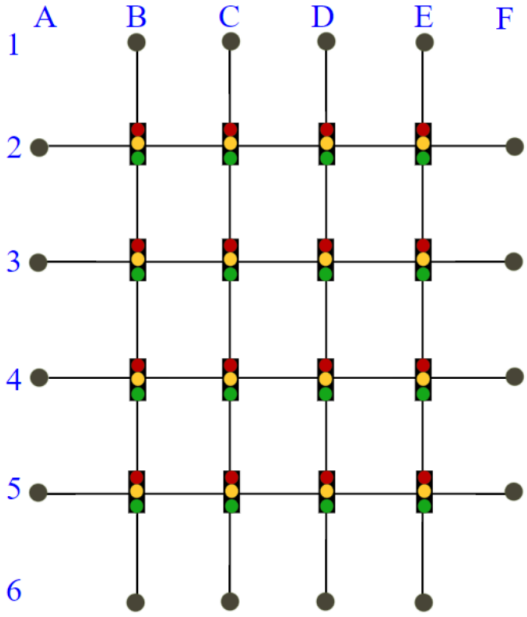
4 实验与结果

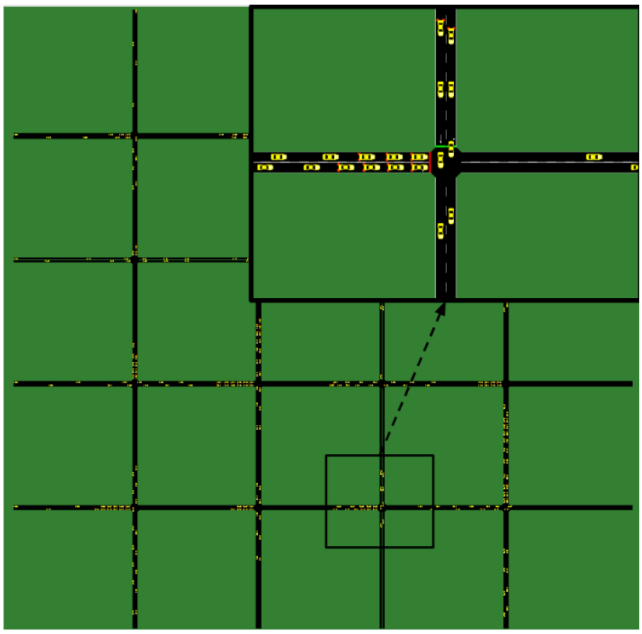
以下实验的主要目标是量化交通信号控制中RL智能体学习过程中不同原因的非平稳性的影响。情境的明显变化(例如，车流量在一个或多个方向上的变化)是这些原因之一，并且在以下所有实验中都存在。本节首先描述正在仿真的场景以及交通情境的详细信息，然后定义所使用的性能指标以及所执行的不同实验。

我们首先进行了一个实验，其中交通信号机使用固定的控制策略-在基础设施中缺乏传感器和/或协调器的情况下的通用策略。本实验的结果将在第4.3节中讨论，并用于强调缺乏可以适应不同背景的策略的问题;它还可以作为以后比较的基准。之后，在第4.4节中，我们将探讨智能体在训练阶段尚未观察到的情境/交通模式中采用给定策略的设置。在第4.5节中，我们分析了(1)当智能体在整个仿真过程中继续探索和更新其Q表时，情境变化的影响;(2)由情境变化和3.1节中介绍的两种不同状态定义的使用引入的非平稳性的影响。然后，在第4.6节中，我们讨论了非平稳性与部分观测之间的关系，这些观测是由观测空间离散性较差的仿真的不精确传感器造成的。最后，在4.7节中，我们讨论了观察到的结果的主要发现和含义。

4.1 场景

我们使用开源微观交通仿真平台SUMO对交通场景及其动态进行建模和仿真，并使用SUMO- rl将仿真实例化为具有MDP所有组件的强化学习环境。交通网络是一个4 × 4的网格网络，所有16个十字路口都有交通信号机。所有连接都有150米，两车道，单向。垂直连接遵循南北向交通方向，水平连接遵循东西向。一共有8对OD: 4对在东西向(A2F2、A3F3、A4F4、A5F5)， 4对在南北向(B1B6、C1C6、D1D6、E1E6)。





为了证明情境变化对交通信号的影响(因此，对交通的影响)，我们定义了两种具有不同车辆流量的不同交通情境。这两个情境每秒在网络中插入相同数量的车辆，但通过在可能的OD对上使用这些车辆的不同分布来实现。特别是:

- 情境1 (NS = WE):在所有8对OD中每3秒插入1辆车的速率。
- 情境2 (NS < WE): N-S方向OD对和1的插入率为每6秒1辆车，车辆每2秒在W-E方向OD对。

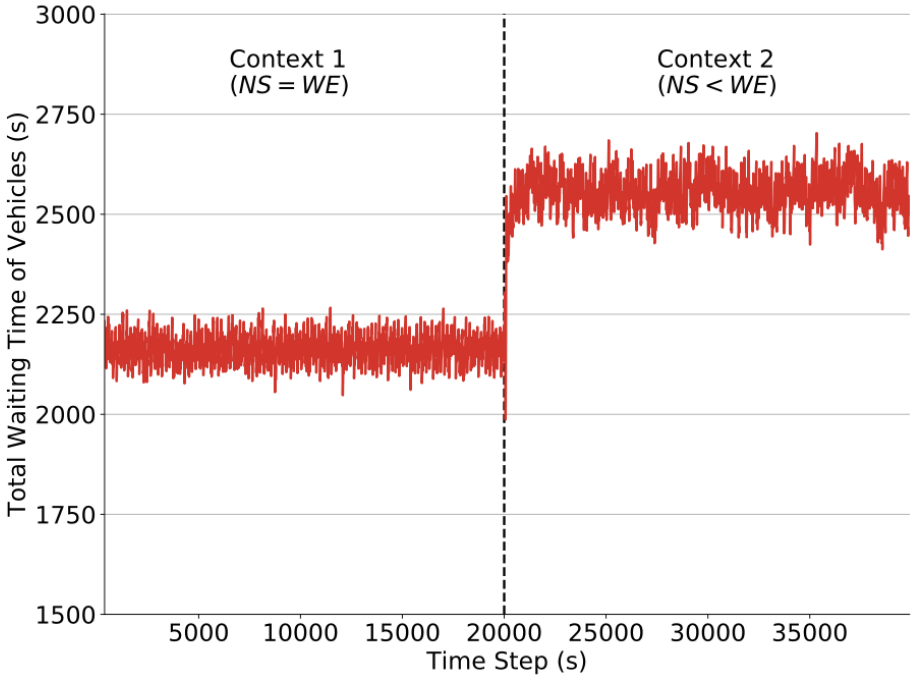
可以预期，两个绿色相位均匀分布的策略在情境1中会有令人满意的表现，但在情境2中则不然。在接下来的实验中，我们每20000个时间步在情境1和情境2之间切换，从情境1开始模拟。这意味着插入速率每20000次改变一次时间步，遵循前面提到的情境。

4.2 度量

为了衡量交通智能体的性能，我们使用所有十字路口累积的车辆等待时间的总和作为度量，如式7所示。直观地说，这量化了由于长时间的等待队列和不充分使用红色相位而不得不将速度降低到0.1米/秒以下的车辆延误的时间。在发生相位变化的时间步长中，由于许多车辆正在停车，许多车辆正在加速，因此队列大小会发生自然振荡。因此，这里显示的所有图都描绘了前面讨论的指标在15秒时间窗口内的移动平均线。与Q-learning相关的图在30次运行中取平均值，其中阴影区域表示标准差。此外，我们省略了模拟开始时的时间步长(因为那时网络还没有完全被车辆填充)以及最后的时间步长(因为那时车辆不再被插入)。

4.3 固定策略交通信号控制

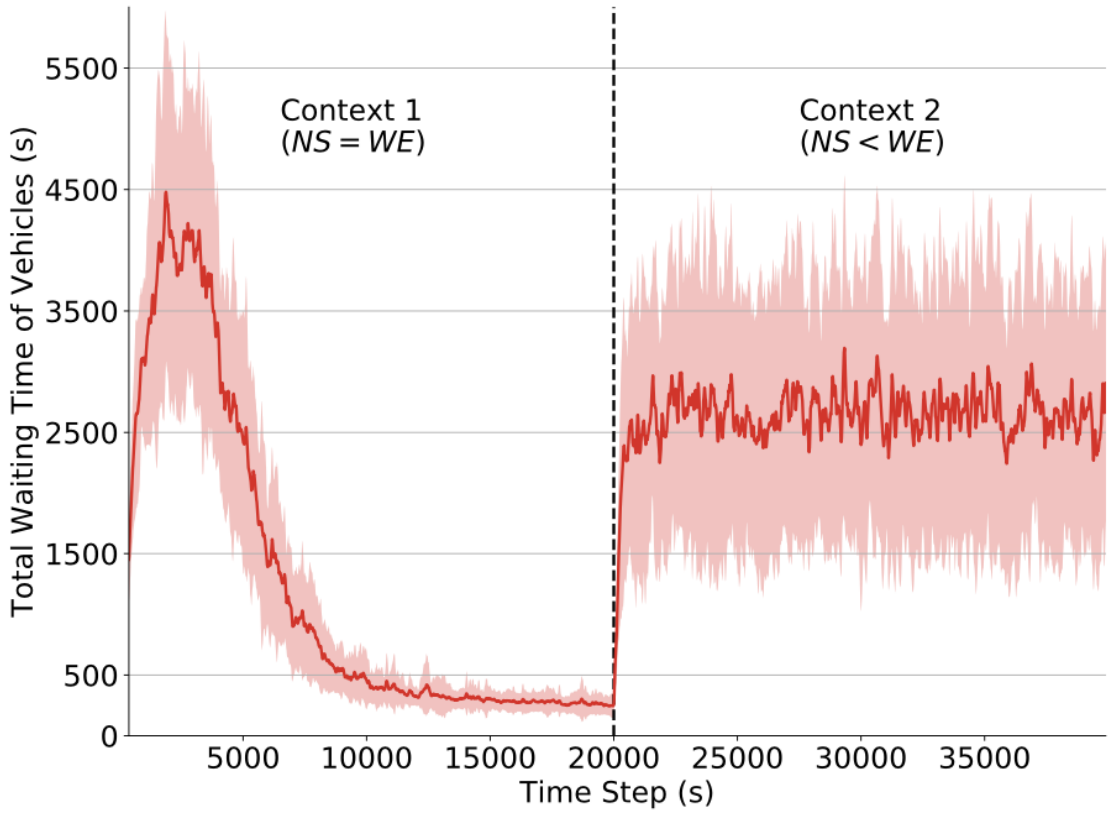
我们首先演示了根据HCM设计的固定策略的性能，该手册通常用于此类任务。固定策略为每个阶段分配35秒的绿灯时间和2秒的黄灯时间。如上所述，我们定义此策略的目标是构建一个基线，用于在两种情况下量化情境变化对交通信号性能的影响：一种情况是交通信号遵循固定策略，另一种情况是交通信号使用QL算法适应和学习新策略。本节分析前一种情况。如图2所示，当情境发生变化时，固定策略会损失性能。当在时间步20000，设置交通流为情境2的时候，更多的车辆在W-E方向行驶，从而产生更大的等待队列。为了使用固定策略获得良好的性能，有必要为每个情境定义一个策略，并提前知道情境发生更改的确切时刻。此外，可能存在任意数量的此类情境，而智能体通常无法事先知道存在多少情境。由于无法预测可能影响环境状态的非重复事件(如交通事故)，因此通常无法获得这些数量的先验知识。因此，在交通流状态可能随时间(缓慢或突然)变化的情况下，固定策略的交通信号控制是不够的。



4.4 禁用学习和探索能力的影响

我们现在描述的情况是，在某个时间点，智能体停止从它们的行为中学习，并简单地遵循给定情境变化之前学到的策略。这里的目标是模拟一种情况，其中交通智能体将先前学习的策略应用于尚未在其训练阶段观察到的情境/交通模式。当情境发生变化时，我们通过将α(学习率)和ε(探索率)设置为0来实现这一点。通过观察Eq. 3，我们看到，如果α = 0, q值不再改变其值。通过设置ε = 0，我们还确保智能体不会探索，并且它们只会选择具有较高估计q值的动作，给定最后观察到的情境的状态。通过分析此设置中的性能，我们可以量化仅通过遵循从以前的情境中学习到的策略而行动的智能体的负面影响。

在训练阶段(直到时间步20000)，我们使用学习率α = 0.1和折现因子γ = 0.99。勘探率从ε = 1开始，每次智能体选择一个动作时，勘探率衰减0.9985。这些定义确保了智能体在开始时大部分都在探索，而到时间步10000时，ε低于0.05，从而导致智能体即使在情境改变后也会继续纯粹地利用当前学习到的策略;也就是说，智能体不能适应上下文的变化。



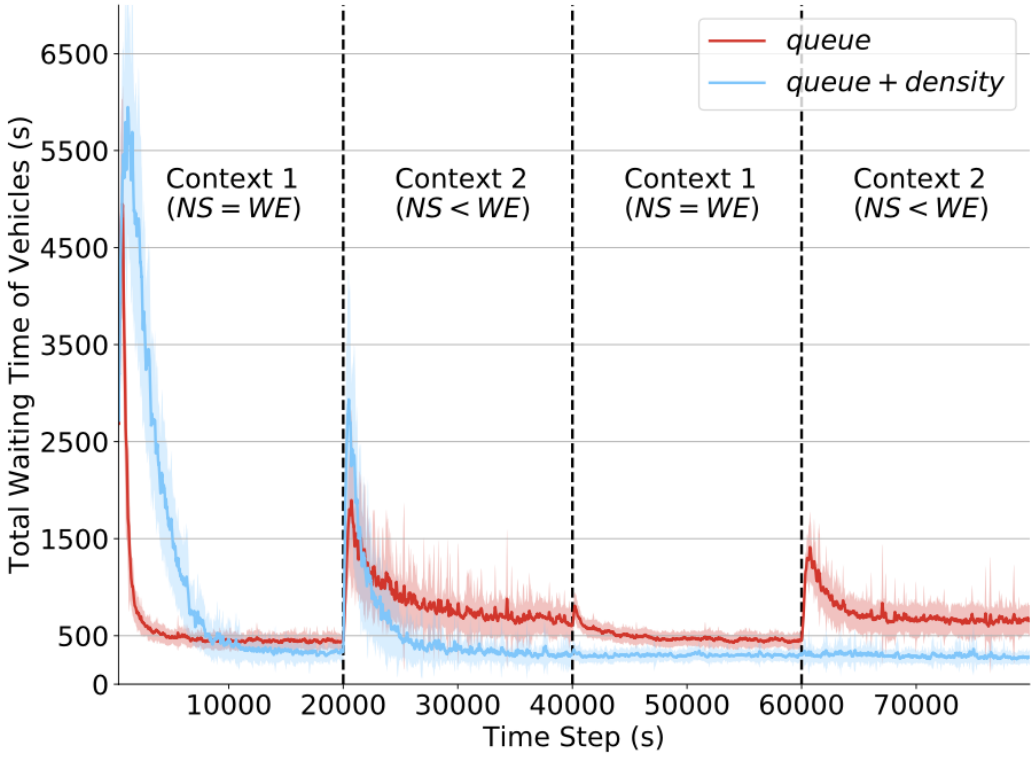
从图3中我们可以看到，在环境变化后(时间步20000)，车辆的总等待时间迅速增加。环境状态的这种变化导致在情境1中学习到的策略不再有效，因为情境2引入了交通信号尚未观察到的交通流量模式。因此，交通信号智能体不知道在这些状态下采取的最佳行动是什么。但是，请注意，某些操作(例如，当其中一个方向出现拥塞时改变相位)仍然能够提高性能，因为它们在两个情境中都是合理的决策。这就解释了为什么当情境发生变化时，性能会大幅下降，以及为什么等待时间在之后一直波动。

4.5 状态可观测性降低的影响

在本实验中，我们比较了3.1节中给出的两种不同状态定义下情境变化的影响。方程4中的状态定义代表了一个更不现实的场景，在这个场景中，昂贵的实时交通传感器可以在十字路口使用。相反，在Eq. 5的部分状态定义中，每个交通信号只有关于在其对应的十字路口(队列)停了多少车辆的信息，但不能将该信息与当前接近其等待队列的车辆数量联系起来，因为运动中的车辆只考虑密度属性。

与之前的实验不同，智能体现在在整个仿真过程中不断探索和更新它们的q表。ε参数设为0.05的固定值;通过这种方式，智能体主要利用但仍然有很小的机会探索其他行动，以

适应环境的变化。通过不改变ε，我们可以确保性能变化不是由勘探策略引起的。QL参数(α和γ)值与上一实验保持一致。



实验结果如图4所示。通过分析仿真的初始步骤，我们注意到使用简化状态定义的智能体学习速度明显快于使用同时包含队列和密度属性的状态定义的代理。这是因为要探索的状态更少，因此策略收敛所需的步骤更少。然而，考虑到这种有限的观察能力，与具有更广泛状态可观察性的智能体相比，智能体收敛到一个策略会导致更高的等待时间。这表明密度属性是更好地表征道路交叉口真实状态的基础。还要注意，在时间步10000左右，两种状态定义的性能(总等待时间约500秒)都优于在固定策略程序下实现的性能(总等待时间约2200秒)，如图2所示。

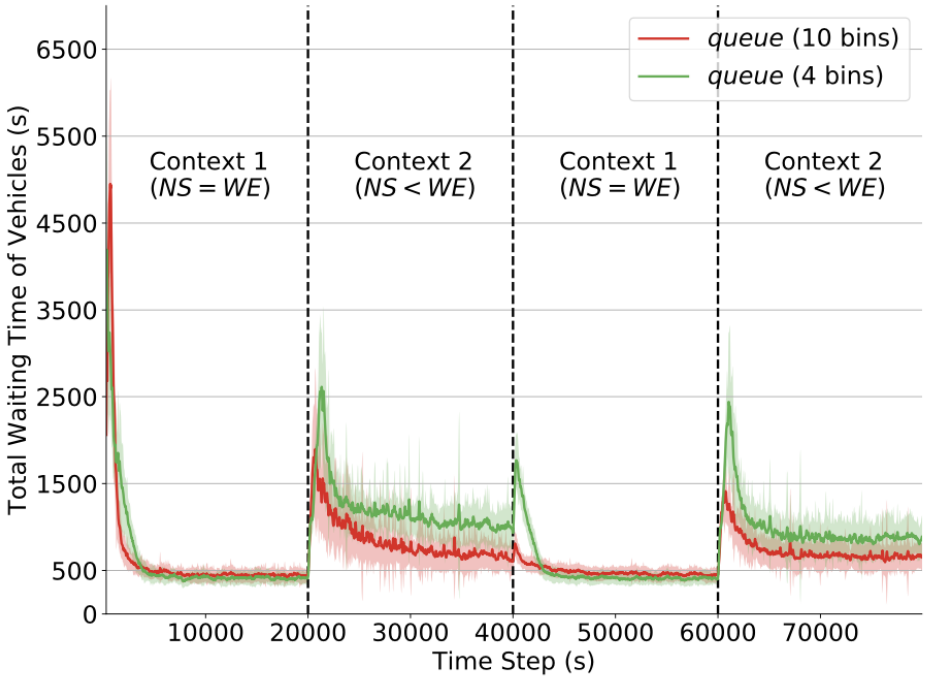
在第一个情境更改中，在时间步20000时，两个状态定义的总等待时间显著增加。这是意料之中的，因为这是智能体第一次必须在情境2中操作。从该情境恢复的在原始状态定义下运行的智能体变化很快，并且获得了与情境1相同的性能。然而，对于部分状态定义(即，只有队列属性)，智能体在情境2下操作时更具有挑战性，情境2描述了到达十字路口的不平衡交通流。

最后，我们可以观察到(在时间步60000处)情境变化引入的非平稳性是如何与有限部分状态定义相关的。同时观察队列和密度的交通智能体在其控制交叉口的等待时间没有振荡，

而只观察队列的交通智能体的性能下降明显。尽管已经经历了情境2，但他们必须重新学习他们的策略，因为过去的q值被学习机制覆盖，以适应不断变化的过去动态。然而，这两种情境的动态在原始状态定义中都得到了很好的捕捉，因为密度和队列属性的组合提供了关于到达十字路口的交通动态的足够信息。这一观察结果强调了更广泛的状态可观察性的重要性，以避免RL代理中非平稳性的负面影响。

4.6 不同层次的状态离散化的影响

除了没有合适的传感器(导致状态描述不完整)，另一个可能导致非平稳性的原因是精度差和观测范围小。举个例子，考虑在十字路口等待车辆数量测量的不精确性;这可能会导致不同的状态（其中不同的行为是最佳的）被认为是相同的状态。这不仅会导致次优性能，而且在情境发生变化时还会导致性能急剧下降。在密度属性不可用的情况下，我们通过降低属性队列离散级别的数量来仿真这种效果。



在图5中，我们描述了当情境发生变化时，属性队列的离散化级别如何影响性能。红线对应于将队列离散为10个均匀分布的级别/bin时的性能(参见3.1节)。绿线对应于离散化水平降低为4个箱时的性能。请注意，在情境更改之后(在时间步20000、40000和60000处)，我们可以观察到减少离散化级别的使用是如何导致性能显著下降的。例如，在40000时刻，在较低离散化水平下运行时，总等待时间增加了3倍。

直观地说，对其真实状态观察不精确的智能体感知转移函数变化的能力会降低。因此，当交通流量在十字路口发生变化时，具有不精确观测的智能体需要更多的动作来重新适应，从而大大增加了队列。

4.7 讨论

已经提出了许多RL算法来解决非平稳问题[15,16,12]。具体来说，这些工作假设环境是非平稳的(没有研究或分析非平稳的具体原因)，然后提出在该设置下有效学习的计算机制。在本文中，我们处理了一个互补问题，即量化学习性能中不同原因的非平稳性的影响。我们也假设存在非平稳性，但我们明确地模拟了许多可能的突出原因，为什么它的影响可能发生。我们研究这个互补问题，因为这是我们的理解，通过明确量化非平稳效应的不同原因，可能会对使用哪种特定算法做出更明智的决定，或者决定，例如，是否应该通过设计更完整的特征集而不是设计更复杂的学习算法来更好地投入精力。

在本文中，我们具体研究了这些可能的原因，当它们影响城市交通环境时。我们的实验结果表明，车辆流量变化形式的非平稳性显著影响遵循固定策略的交通信号控制器和从标准RL方法中学习的策略，这些策略不适用于不同的情境。然而，这种影响(导致路口等待车辆总数的快速变化)对智能体的影响程度不同，取决于这些智能体可观察性的不同程度。虽然具有原始状态定义(队列和密度属性)的智能体只在它们第一次在新情境中操作时表现出性能下降，但具有较少观察值(仅队列属性)的智能体可能总是必须重新学习重新适应的q值。然而，最初的状态定义在现实世界中不是很现实，因为能够为大型交通道路提供这两种属性的传感器非常昂贵。最后，在智能体只观察队列属性的情况下，我们证明了不精确的度量(例如，低离散箱数)可能会导致情境变化的影响。因此，为了设计一个鲁棒的RL交通信号控制器，关键是要考虑哪些是最合适的传感器，以及它们如何有助于提供对真实环境状态的更广泛观察。

我们观察到，在竞争环境中，由其他并发学习智能体的行为引入的非平稳性似乎是获取有效交通信号策略的一个小障碍。然而，一个自私地学习减少自己队列大小的交通智能体可能会引入更高的车辆流量到达邻近的十字路口，从而影响其他代理的奖励并产生非平稳性。我们相信，在更复杂的情况下，这种影响将更加明显。
此外，我们发现，如果不考虑非平稳性影响，传统的表格独立q学习在我们的场景中表现良好。因此，在这个特殊的仿真中，没有必要使用更复杂的方法，如基于值函数近似的算法;例如，深度神经网络。这些方法可以帮助处理可能需要处理高维状态的大规模仿真。然而，我们强调的事实是，尽管它们可以帮助处理高维状态，但它们也会受到非平稳性的影响，就像标准表格方法一样。这是因为就像标准的表格q学习一样，深度强化学习方法没有明确地对非平稳性的可能来源进行建模，因此每当状态转移函数发生变化时，学习性能就会受到影响。

5 相关研究

强化学习在以前已经成功地用于交通信号控制的解决方案中。该领域的调查[17,18,19]讨论了交通信号控制强化学习的基本方面，如状态定义、奖励函数和算法分类。在这种情况下，许多研究已经解决了多智能体RL[20,6,8]和深度RL[21,22,23]方法。尽管非平稳性经常被认为是交通领域的一个复杂挑战，但我们证明缺乏量化其影响并将其与许多原因和结果联系起来的工作。

在表1中，我们比较了以部分可观测性、车辆流量分布变化和/或多智能体场景的形式解决非平稳性的相关工作。da Silva等人在[12]中探讨了不同交通模式下交通信号控制的非平稳性。他们提出了RL-CD方法来创建环境的部分模型——每个模型负责处理一种上下文。然而，他们使用了一个简单的状态和每个交通信号代理可用的动作模型:状态被定义为每个传入链路的占用，并离散为3个bin;动作包括从三个固定的和预先设计的信号方案中选择一个。Oliveira等人在[24]中扩展了[12]的工作，解决了驾驶员在驾驶操作任务(如减速概率)方面的随机行为所导致的非平稳性问题，但没有改变上述简单的状态和动作模型。在[23]中，Liu等人提出了一种独立深度q学习的变体来协调四个交通信号。然而，没有提到或分析有关车辆分布或插入率的信息。[7]比较了使用A3C算法的不同状态表示;然而，该论文没有研究智能体适应不同交通流分布的能力。[25]在车辆到基础设施(V2I)场景中分析了状态可观察性，其中交通信号代理以不同速率检测使用专用短程通信(DSRC)技术的接近车辆。在[26]中，研究了部分可观察状态的场景(只有占用传感器可用)，但没有对不同状态定义或传感器进行比较。在[9]中，Chu等人在网络中分布的不同车辆流独立改变其插入率的场景下引入了Multiagent A2C。另一方面，他们只使用了一个状态定义，它给出了关于交通路口的足够信息。最后，在[27]中，Padakandla等人引入了context - ql，这是一种类似于RL-CD的方法，使用变更点检测度量来捕获上下文变化。他们还探讨了由不同交通流量引起的非平稳性，但他们没有考虑在他们的结果中使用的状态定义(具有低离散化和只有一个传感器)的影响。就我们所知，这是第一次分析在非平稳环境下，不同程度的部分可观测性如何影响交通信号代理，在这种环境下，交通流不仅在车辆插入率上发生变化，而且在不同阶段之间的车辆插入分布上也发生变化。

6 结论

非平稳性是将强化学习应用于现实世界问题的一个重要挑战，特别是交通信号控制。在本文中，我们研究并量化了不同原因对学习代理性能的非平稳性的影响。具体来说，我们研究了多智能体交通信号控制中的非平稳性问题，其中非平稳性是由交通模式的显式变化和状态可观察性降低引起的。这种类型的分析补充了与RL非平稳性相关的现有工作;这些研究通常提出在不断变化的环境下学习的计算机制，但通常没有系统地研究不同来源的非平稳性可能对学习绩效产生的具体原因和影响。

我们已经证明，独立的Q-Learning代理可以根据交通模式上下文的变化重新调整策略。此外，我们已经证明，智能体的状态定义和它们的观察范围强烈影响智能体的再适应能力。虽然具有更广泛状态可观察性的智能体在动态变化到以前经历过的上下文时不会经历性能下降，但在状态的部分可观察版本下运行的智能体通常必须重新学习策略。因此，我们已经证明了如何更好地理解非平稳性的原因和影响可能有助于RL代理的开发。特别是，我们的研究结果表明，与设计更复杂的学习算法相比，设计更好的传感器和状态特征可能对学习性能有更大的影响。