智能交通系统中的交通信号灯控制深度强化学习

Xiao-Yang Liu[+]，*IEEE*会员，Ming Zhu[+*]，*IEEE*会员，Sem Borst，*IEEE*会员，以及 Anwar Walid，*IEEE*会士

arXiv:2302.03669v2 [cs.LG] 5 Mar 2023

摘要—**智能交通系统（ITS）中的智能交通信号灯有望显著提高交通效率并缓解交通拥堵。深度强化学习（DRL）是一种很有前景的方法，可根据道路网络中的实时交通状况自适应地控制交通信号灯。然而，传统方法可能存在可扩展性较差的问题。本文研究了利用深度强化学习控制交通信号灯；理论分析和数值实验均表明，在网格道路网络中，智能行为“绿波”（即车辆将连续遇到绿灯，无需在任何交叉口刹车）会自然形成；并证明，在一条设有多个横向街道的大道上，这种行为是最优策略。作为第一步，我们在两种场景中采用两种深度强化学习算法解决交通信号灯控制问题。在单个道路交叉口，我们验证了深度Q网络（DQN）算法能够产生阈值策略；在网格道路网络中，则采用深度确定性策略梯度（DDPG）算法。其次，数值实验表明，DQN算法能够实现最优控制；而采用**被动观测**的DDPG算法能够自行在网格道路网络中形成高层智能行为，即涌现出“绿波”策略。我们还在一个 5×10 网格道路网络中验证了“绿波”模式。第三，“绿波”模式表明，深度强化学习算法能够产生良好的解，因为实验结果所呈现的“绿波”策略已被证明是在特定交通模型（设有多个横向街道的大道）中的最优策略。在单个道路交叉口和网格道路网络中得到的策略均表明，深度强化学习算法具有良好的可扩展性。**

关键词—**交通信号灯控制，智能交通系统，马尔可夫决策过程（MDP），深度强化学习，可扩展性，绿波，高层智能行为。**

I. 引言

智能交通信号灯在智能交通系统（ITS）中发挥着重要作用 [1]–[4]，因为它们可以显著提高交通效率并缓解交通拥堵。根据一份报告 [5]，，美国因交通拥堵造成的年度成本已从2000年的750亿美元增至2017年的1790亿美元。道路交叉口处车辆的能源消耗和排放与交通信号灯的控制策略密切相关 [6]。深度强化学习（DRL） [7]–[9] 是一种很有前景的方法，可根据实时交通状况（例如各条道路上的交通流量）自适应地调整交通信号灯的控制策略。

传统交通信号灯控制方法可能存在可扩展性差的问题。1）基于经典优化的方法[10]–[19] 根据交通数据计算出的参数以及基于专家知识制定的启发式规则作出决策。如果道路网络规模较大，参数数量也会很多，需要大量计算资源，导致可扩展性差。这些启发式规则基于经验，可能无法涵盖复杂交通场景中的所有情况，缺乏灵活性。例如，在 Maxband [12]，中，所有交叉口必须采用相同的周期长度，即所有交叉口中的最大周期长度；在 Max-pressue [14]，中，只有当当前相位时间大于根据专家知识设定的最小相位持续时间时，才会计算压力（进口道队列总长度减去出口道队列总长度）。2）自适应交通信号灯控制问题可以表述为马尔可夫决策过程（MDP） [20]–[22]，，因为控制动作的影响仅取决于当前状态，而与先前历史无关。值迭代、策略迭代和动态规划等传统的基于MDP的方法能够针对相对小规模的问题求得最优策略。然而，状态-动作空间会随道路交叉口数量增加而呈指数级增长，在大型道路网络中将变得不可行。3）一些深度强化学习算法（例如深度Q网络（DQN）算法）也不具备可扩展性。DQN算法中的动作是离散的。道路交叉口数量增加时[8]，[23]，，待估计的Q值数量会呈指数级增长。实际中无法近似如此大量的Q值。因此，在大型道路网络中，一些深度强化学习算法（例如DQN）可能无法产生可行解。

深度强化学习算法 [24]，[25] 在交通信号灯控制中可能无法获得理想的解决方案。神经网络用于近似 Q 函数 [26] 和/或策略；然而，它们是黑箱，且可解释性较差。神经网络接收状态（例如每条道路上的等待车辆数量以及交通信号灯配置）和动作（例如“继续”或“切换”），并输出估计的 Q 值，却不提供任何解释。无法保证 DRL 会收敛到某种随机梯度下降方法。换言之，无法保证输出动作在某一状态下所有动作中产生最大的 Q 值。此外，还可能出现过拟合或欠拟合

[+]贡献相同。
[*]通讯作者。X.-Y. Liu 任职于美国纽约 10027 哥伦比亚大学电气工程系。电子邮箱：{xl2427}@columbia.edu。M. Zhu 的研究工作得到中国国家自然科学基金（项目编号：61902387）资助。M. Zhu 任职于中国科学院深圳先进技术研究院，中国深圳 518055。电子邮箱：zhumingpassional@gmail.com。S. Borst 和 A. Walid 任职于诺基亚贝尔实验室，电子邮箱：sem.borst@outlook.com、anwar.walid@nokia-bell-labs.com。

。因此，估计的 Q 值可能不准确，输出动作也可能无法产生最大的 Q 值。

本文研究了用于控制交通信号灯的深度强化学习方法，并分析了其性能。在单个道路交叉口中，我们证明，DQN算法得到的策略与传统MDP方法得到的最优控制相匹配。在网格道路网络中，我们采用深度确定性策略梯度（DDPG）算法；由于状态-动作空间过大，通过传统MDP方法获得最优策略并不实际；我们还证明，在指定的交通模型中，智能行为“绿波”（即车辆将一路遇到连续绿灯，无需在任何交叉口刹车）在理论上是最优的。数值结果表明，应用DDPG算法可以产生“绿波”策略，证实了该算法能够实现高层智能行为。

我们的主要贡献可总结如下：

- 我们在两种场景下使用两种深度强化学习算法来解决交通信号灯控制问题。在单个道路交叉口中，我们验证了深度Q网络（DQN）算法能够产生与最优控制相匹配的阈值策略；在网格道路网络中，我们采用深度确定性策略梯度（DDPG）算法。
- 数值实验表明，采用被动观测 的DDPG算法能够自主地在网格道路网络中产生高层次的智能行为，即涌现出“绿波”策略。我们还研究了“绿波”模式。
- DDPG算法产生的“绿波”模式表明，深度强化学习算法能够得出良好的解，因为在特定的交通模型（含有多条横向街道的大道）中，“绿波”策略是最优的。单个道路交叉口和网格道路网络中的策略都展现了深度强化学习算法在解决交通信号灯控制问题方面的可扩展性。

本文其余部分安排如下。第二节介绍相关工作。第三节详细描述道路交叉口模型并陈述问题。第四节说明适用于单个交叉口和网格道路网络的深度强化学习算法。第五节评估所提出的深度强化学习算法的性能，并展示“绿波”模式的涌现。第六节总结全文。

II. 相关工作

现有的交通信号灯控制研究可分为两类：传统方法和深度强化学习方法。

A. 传统方法

交通信号灯控制实际上是一个混合整数线性规划问题 [27] ，属于 NP难问题。交通信号灯控制高度依赖准确的交通参数，例如车辆的位置和速度。传统技术使用传感器 [28]–[30], 测量这些参数，例如红外传感器和感应线圈检测器。为了实时收集交通数据并为大量交叉口计算策略，需要较高的数据传输速率和充足的计算资源。无线通信 [31], [32] 和边缘云计算 [33]等技术可以显著提高数据传输速率和计算效率。传统方法主要包括经典优化方法[10]–[19],、进化算法 [34]–[36], 和模糊逻辑[37], [38]。

经典优化方法 [10]–[19] 通常根据参数和启发式规则调整周期时间和相位分配。绿灯时间通常长于实际所需，导致等待车辆或行人浪费大量时间。

- 单个道路交叉口的方法。Webster方法先计算所需的周期长度，再计算绿灯时间分配，其分配比例取决于各相位关键车道流量的比值。Hofri等人 [19] 针对有两条队列的单个道路交叉口提出了一种阈值策略，该策略依赖交通流数据和仿真实验；然而，它可能无法适用于所有实际交通状况。
- 多道路口的方法。GreenWave[11]和 Maxband [12] 方法都旨在减少车辆停车次数；然而，GreenWave仅优化一个方向，而Maxband优化两个相反方向。GreenWave和Maxband要求所有交叉口采用相同的周期长度。在自组织交通信号灯（SOTL） [13] 方法中，会测量当前相位和其他竞争相位对绿灯信号的请求，然后制定特定规则来决定“继续”还是“切换”。其性能高度依赖这些规则。在Max-pressure [14] 方法中，会计算所有相位的压力，然后选择压力最大的相位。在悉尼协调自适应交通系统（SCATS） [15] 方法中，会计算当前信号方案中各相位的饱和度（DS），即有效绿灯时间与可用绿灯时间之比；随后计算其他信号方案的饱和度，最后选取总饱和度最低的方案。为减少交通流的大幅波动，Lammer等人 [17] 提出了一种调整绿灯相位持续时间和顺序的方法，以实现队列和红灯持续时间的稳定化。

这些方法的可扩展性较差。道路网络规模较大时，参数数量也会随之增加，导致参数计算和调优耗时很长。此外，启发式规则和某些参数高度依赖专家知识，且无法保证最优性。

进化算法 [39], [40] ，尤其是遗传算法，通常用于多目标优化，例如燃料消耗、停车次数、交通流吞吐量和车辆等待时间 [34]–[36]。从理论上讲，这些算法基于适者生存的概念，通过随机优化和启发式方法进行求解。以遗传算法为例，解被编码为染色体，在下一代中通过繁殖、

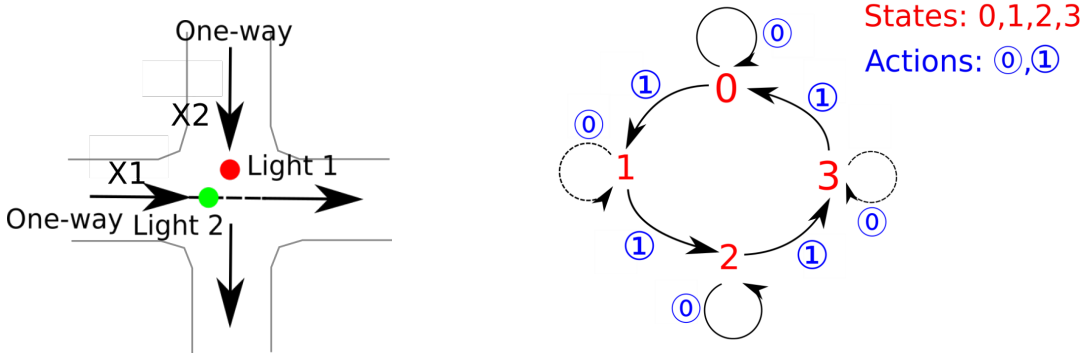

图 1. 一个包含两股交通流的交叉口（左），以及对应的状态转移图（右）。

交叉和极低概率的变异生成新的子代，适应度值较高的个体得以存活。如此一来，在每一代或每次迭代中，解都会逐渐接近局部甚至全局最优点。Brian 等人 [36] 提出了一种遗传算法，用于最大限度地降低燃料消耗、排放、延误和停车次数。Turky 等人 [35] 使用遗传算法动态控制红绿灯时长，以优化车辆和行人流量吞吐量。然而，遗传算法的性能可能取决于种群规模、代数和变异概率等参数，而且没有性能保证。例如，如果种群规模较小，遗传算法可能无法找到有效解。因此，在有限时间内，进化算法可能得到不理想的解。

模糊逻辑 [37], [38] 可用于交通信号灯控制问题，因为语言术语（例如低速、中速和高速）中的模糊性可以用模糊集表示。为补偿交通流波动，Ali 等 [37]使用模糊逻辑监测并处理连续两个周期之间交通状况的变化。Hawi等 [38] 结合模糊逻辑和无线传感器网络来控制交通信号灯，其中无线传感器网络实时收集交通数据，模糊逻辑则以交通量和等待时间为输入，输出绿灯分配顺序。在上述方法中，设计模糊逻辑需要较高的专业知识，并且需要定期更新规则，而且通常无法给出准确的推理。此外，其性能取决于启发式规则和参数，且无法保证稳定性或最优性。

B. 深度强化学习方法

深度强化学习方法有多种构建方式，包括状态、动作和奖励，因此会产生不同的结果。根据这些构建方式应用深度强化学习算法，并将这些算法分为单智能体和多智能体两类。

首先，我们讨论深度强化学习的构建方式。

- 状态表述。状态通常是以下项目的组合：队列长度 [41], 等待时间[42],交通量 [43] （车道/道路上的车辆数量）、延误 [44]（实际行驶时间减去预期行驶时间）、速度、相位时长（当前相位的时长） [45], 交通拥堵 [46]（以 0/1 或交通拥堵程度表示），以及图像或矩阵 [47], [48] （展示车辆的位置，其中“0”表示不存在，“1”表示存在）等。近年来，越来越多地使用图像/矩阵作为状态；然而，状态的维度非常大，例如超过一千维，这会导致较长的训练时间。
- 动作构建。动作通常有四种类型：设置当前相位时长 [49]，设置相位比例 [50],继续/切换 [51], 以及选择下一个相位 [52]。前两种动作是连续的，后两种是离散的。选择下一个相位更加灵活，因为交通信号灯无需循环改变。这里我们选择继续/切换，因为它具有综合性。
- 奖励构建。奖励通常包括以下项目：队列长度 [41], 等待时间 [51],通行量 [49], 速度 [51], 以及压力 [53]（驶入车辆数减去驶出车辆数）。一些研究将上述项目进行加权线性组合，作为奖励。权重调整是一个重要问题，权重的微小变化可能导致结果大不相同。为避免这种情况，我们仅选择队列长度作为奖励。

其次，我们介绍单智能体和多智能体深度强化学习方法。在单智能体深度强化学习方法中，智能体学习每个状态下所有动作的Q值，并选择该状态下近似Q值最大的动作。通过自举学习过程，智能体利用原始输入数据为端到端任务获得策略。由于可扩展性较差，这些方法仅适用于小型或中型道路网络。深度Q网络及其变体 [8], [25] 是常用方法。Liang 等人 [8] 提出使用决斗双重深度Q网络和优先经验回放来确定交通信号灯时长。单智能体深度强化学习方法能够全局掌握交通状况；但在超大型道路网络中可能并不实用。

在多智能体深度强化学习方法 [42], [54]–[58], 中，每个智能体控制一个交叉口的交通信号灯，同时考虑其邻域的状态和动作，以便与邻域协作。邻居智能体的加权奖励之和加上自身奖励，

随后被视为全局奖励。Liu等人 [56] 提出了一种带函数近似的协作式Q学习算法，使每个局部交叉口都能在没有中央监督智能体的情况下协同决策。Chu等人 [42]提出了一种基于行动者-评论家的多智能体深度强化学习方法，以稳定学习过程、提高可观测性并降低学习难度。Xu 等人 [55] 提出了一种多智能体深度强化学习方法，用于优化交叉口的信号配时方案。Wu 等人 [54]提出了一种用于交通信号灯控制的多智能体深度确定性策略梯度方法，其中采用长短期记忆网络（LSTM） [59] 来改善部分可观测状态导致的环境稳定性问题。多智能体深度强化学习方法能够让智能体在超大规模道路网络中以协作方式控制交通信号灯。然而，这些方法通常可能无法全面掌握交通状况，而这对于制定良好的控制策略至关重要。此外，如何平衡智能体及其邻域的效用也是一个难题。

上述深度强化学习算法使用黑箱神经网络拟合 Q 函数，并采用随机梯度下降方法最小化非凸损失函数。这些方法无法保证可扩展性。在深度Q网络方法中，每个状态下所有动作对应的近似 Q 值数量会随着道路交叉口数量的增加而呈指数增长。DQN算法需要大量计算资源和时间来估计大型道路网络中的 Q 值，导致可扩展性较差。多智能体深度强化学习方法，例如多智能体深度确定性策略梯度 [54] ，也可能无法保证可扩展性，因为它们是黑箱模型，且可能无法获得最优策略。

不同于现有方法，我们分别将深度强化学习方法，即深度Q网络和深度确定性策略梯度，应用于优化单个道路交叉口和网格道路网络中的交通信号灯控制策略。考虑到DQN算法的可扩展性较差，我们仅将其用于单个道路交叉口。考虑到DDPG算法能够利用连续动作处理复杂场景，我们将其用于网格道路网络。我们使用传统方法获得单个道路交叉口中的最优策略，并在指定交通模型（即“绿波”策略）下获得网格道路网络中的最优策略。仿真结果表明，DQN算法能够实现最优控制，而所提出的DDPG算法能够实时产生高层智能行为，即呈现出“绿波”模式。

III. 模型描述与问题表述

本节介绍单个和多个交通信号灯的模型及优化目标。为便于阐述，我们采用一个经过简化的模型，尽管如此，该模型仍涵盖了决定道路交叉口交通流动态的最关键特征。我们的核心目标是广泛探究深度强化学习算法在优化实时控制策略方面的性能和可扩展性，而非为某个特定实例开发实用策略。因此，我们采用离散时间表述以简化说明，并便于使用传统的马尔可夫决策过程技术进行比较。这些方法和结果可以自然地扩展到连续时间运行。

A. 单个交通信号灯模型

我们首先研究单交叉口场景，以便验证深度强化学习算法。我们考虑一个最简单且有意义的设置，其中两股相交的单向交通流如图 1 左侧示意图所示。我们假设有两个方向，分别用 1 和 2 表示由西向东的方向（大道）和由北向南的方向（横向街道）。因此，共有两股交通流，分别记为 1 和 2。时隙 t 开始时的状态 $S(t)$ 可用三元组 $(X_1(t)\ X_2(t);\ L(t))$ 表示，其中 $X_i(t)$ $i \in \{1,2\}$ 表示交通流 i 中等待通过交叉口的车辆数量，而 $L(t) \in \{0,1,2,3\}$ 表示交通信号灯的配置：

- “0”，即“绿灯”：车流 1（大道）亮绿灯，因此车流 2（横向街道）亮红灯；
- “1”，即“黄灯”：车流 1（大道）亮黄灯，因此车流 2（横向街道）亮红灯；
- “2”，即“红灯”：车流 1（大道）亮红灯，因此车流 2（横向街道）亮绿灯；
- “3”，即“橙色”：车流 1（大道）亮红灯，因此车流 2（横向街道）亮黄灯。

每种配置 $L(t)$ 都可以在下一个时隙保持不变，也可以切换到自然的后续配置 $(L(t)+1) \bmod 4$。这由时隙 t 结束时选择的动作 $A(t) \in \{0,1\}$ 决定，该动作用二进制变量表示：“0”表示保持不变，“1”表示切换。于是有

$$L(t+1) = (L(t) + A(t)) \mod 4. \tag{1}$$

这些规则形成了严格的循环控制序列，如图 1 右侧所示。

队列状态随时间的演变由以下递推关系决定

$$\begin{aligned}&(X_1(t+1), X_2(t+1))\\ &= (X_1(t) + C_1(t) - D_1(t), X_2(t) + C_2(t) - D_2(t)),\end{aligned} \tag{2}$$

其中，$C_i(t)$ 表示交通流 i 在时隙 t 到达交叉口的车辆数量，$D_i(t)$ 表示交通流 i 在时隙 t 穿过交叉口的离开车辆数量。我们作出简化假设：如果某交通流获得绿灯，则在该时隙内恰有一辆等待车辆（如果有）穿过交叉口，即

$$\begin{aligned}&D_1(t) = \min\{1, X_1(t)\}, && \text{if } L(t) = 0,\\ &D_1(t) = 0, && \text{if } L(t) \neq 0;\\ &D_2(t) = \min\{1, X_2(t)\}, && \text{if } L(t) = 2,\\ &D_2(t) = 0, && \text{if } L(t) \neq 2.\end{aligned} \tag{3}$$

B. 多个交通信号灯模型

为考察深度强化学习算法在更复杂场景中的可扩展性，我们考虑如图 2 所示的网格道路网络。具体而言，我们研究一个包含 N 个交叉口和双向交通流的网格道路网络，代表一条设有多条横向街道的大道。我们不考虑任何进行左转或右转的交通流，

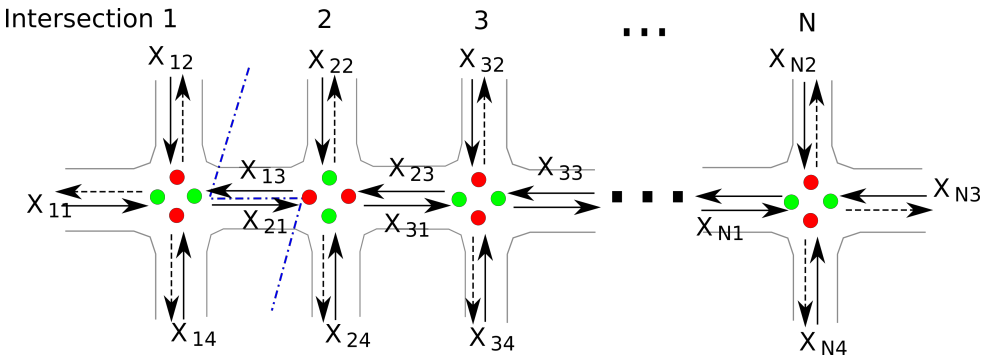

图2. 具有网格拓扑的道路网络示意图。

因为该模型可以轻松推广以适应这种情况。时隙 t 开始时的状态 $S(t)$可表示为一个 $(5N)$ 元组（$X_{n1}(t)$ $X_{n2}(t)$ $X_{n3}(t)$ $X_{n4}(t)$; $L_n(t))^{n\in\{1,\ldots,N\}}$。前四个分量表示等待通过第 n个交叉口的车辆数量，其中 1、2、3 和 4 分别表示由西向东、由北向南、由东向西和由南向北的方向。与之前类似，最后一个分量 $L_n(t)$ $\in\{0,1,2,3\}$ 表示第 n个交叉口的交通信号灯配置：

$$L_n(t+1) = (L_n(t) + A_n(t)) \mod 4, \tag{4}$$

其中， $A_n(t)$表示在时隙 t结束时为第 n个道路交叉口选择的动作。各种队列状态的演变由递推关系决定

$$X_{ni}(t+1) = X_{ni}(t) + C_{ni}(t) - D_{ni}(t), \tag{5}$$

其中， $C_{ni}(t)$表示在时隙 t 期间出现在第 n个交叉口、方向为 i的车辆数量， $D_{ni}(t)$表示在时隙 t期间沿方向 i 通过第 n个交叉口的车辆数量， $i\in\{1,\ldots,4\}$， $n\in\{1,\ldots,N\}$。注意，

$$\begin{aligned}
&D_{n1}(t) = \min\{1, X_{n1}(t)\}, \quad D_{n3}(t) = \min\{1, X_{n3}(t)\}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if } L_n(t) = 0, \\
&D_{n1}(t) = 0, \quad D_{n3}(t) = 0, \quad \text{if } L_n(t) \neq 0; \\
&D_{n2}(t) = \min\{1, X_{n2}(t)\}, \quad D_{n4}(t) = \min\{1, X_{n4}(t)\}, \\
&\qquad\qquad\qquad\qquad\qquad\qquad \text{if } L_n(t) = 2, \\
&D_{n2}(t) = 0, \quad D_{n4}(t) = 0, \quad \text{if } L_n(t) \neq 2.
\end{aligned} \tag{6}$$

C. 优化目标函数

我们假设时隙 t中的“拥堵成本”可表示为队列状态的函数。采用二次函数 $Z(X(t))$ 来表示拥堵成本。在单个道路交叉口场景中，$X(t)$ $=(X_1(t)$ $X_2(t))$，且 $Z(X(t))$ $= X_1^2(t) + X_2^2(t)$。在包含 N 个交叉口的网格道路网络中， $X(t)$ $=$ $(X_{n1}\ t\ \ X_{n2}\ t\ \ X_{n3}\ t\ \ X_{n4}\ t\ \ _{n\in\{1,\ldots,N\}}\ \ Z\ X\ t\ \ =(\,)\,(\,)\,(\,)\,(\,))$，且 $(\,(\,))\,(\,)$。$\sum_{n=1}^{N}\sum_{i=1}^{4} X_{ni}^2\ t$

优化目标是在时隙 $\{0,1,\ldots,T\}$ 内选择动作，以最小化长期期望折扣成本 $\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t Z(X(t))\right]$，其中 $\gamma\in(0,1]$ 表示折扣因子。

IV. 深度强化学习算法

本节首先简要介绍强化学习和Q学习。随后，我们针对不同场景提出两种深度强化学习方案：DQN [25] 和 DDPG [60] 算法。我们在图1所示的单个道路交叉口中使用DQN算法，在图2所示的网格道路网络中使用DDPG算法。DDPG算法在连续动作空间中采用演员-评论家框架。我们的实验中，交通信号灯的动作是离散的，即 $A(t)$ $\in\{0,1\}$，因此我们对传统的DDPG算法采用离散化过程。最后，我们证明，在网格道路网络中，“绿波”控制策略是最优的。

A. 强化学习与Q学习

成本函数的定义。我们将策略 π 下道路交叉口在时隙 $\{0,1,\ldots,T\}$ 上的折扣拥堵成本定义为，其中折扣因子为 γ

$$V_\pi = -\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t|X(t)|^2\right], \tag{7}$$

其中，在单个交叉口或网格道路网络中， $|X(t)|^2$ 定义 $= |X_1(t)|^2 + |X_2(t)|^2$ 。III-C小节中的优化目标是最小化长期期望队列长度，因此我们取其负值，以与最大化操作保持一致。最优策略 $\hat{\pi}$ 满足

$$V_{\hat{\pi}} = -\sup\frac{1}{T}\mathbb{E}\left[\sum_{t=0}^{T}\gamma^t|X(t)|^2\right]. \tag{8}$$

算法1：单个道路交叉口或网格道路网络的一步环境仿真

1：根据算法 2 [中第 3 行，在单个道路交叉口] 或网格道路网络]中选择动作 a ； 2：在单个道路交叉口] 中生成随机变量 C_1, C_2 [，或在网格道路网络]中为所有 $n \in \{1, ..., N\}$[生成随机变量 C_{n1}, C_{n3} 和 C_{n2}, C_{n4}； 3：执行动作 a，根据公式 (1) ∼–(3)[确定新的队列状态和交通信号灯状态 X_1', X_2', L' （适用于单个道路交叉口]），或根据公式 (4) ∼–(6)[为所有 $n \in \{1, ..., N\}$ 确定相应状态$X_{n1}', X_{n2}', X_{n3}', X_{n4}' = 0$（适用于网格道路网络]）； 4：计算单个道路交叉口] 的奖励 $r = -(|X_1'|^2 + |X_2'|^2)$，或计算网格道路网络]的奖励 $r = -\sum_{n=1}^{N}\sum_{i=1}^{4}|X_{ni}'|^2$ [； 5：更新状态：在单个道路交叉口] 中更新 $X_1 = X_1', X_2 = X_2', L = L'$ [，或在网格道路网络]中为所有 $n \in \{1, ..., N\}$ [更新 $X_{n1} = X_{n1}', X_{n2} = X_{n2}', X_{n3} = X_{n3}', X_{n4} = X_{n4}', L_n = L_n'$。

算法2：单个道路交叉口的深度Q网络

1：使用随机权重 θ_0初始化 $X_1 = X_2 = 0$、 $L = 0$和动作价值 Q，设置批大小 M、时间步数 K以及经验回放缓冲区 D； 2：对于步数 $k = 1$，…， K3：以概率 $1 - \epsilon$选择 $a = \arg\max_{\bar{a} \in \mathcal{A}} Q(s, \bar{a})$，或以概率 ϵ选择随机动作 a； 4：执行动作 a，获取新状态 $s' = [X_1', X_2'; L']$并从算法1获取奖励 r； 5：将状态转移(s, a, r, s')存入经验回放缓冲区 D； 6：从 {个状态转移中抽取一个包含 M个样本的小批量}$_j(s_t^j, a_t^j, r_t^j, s_{t+1}^j)$； 7：使用Adam优化器更新网络：$\nabla_{\theta_k}$损失$(\theta_k) = \mathbb{E}_{s_t^j, a_t^j, r_t^j, s_{t+1}^j}\Big[(r + \gamma \max_{a_{t+1}^j}\big(Q(s_{t+1}^j, a_{t+1}^j|\theta_{k-1})\big) - Q(s_t^j, a_t^j|\theta_k))\nabla_{\theta_k} Q(s_t^j, a_t^j|\theta_k)\Big]$。

算法3：网格道路网络的深度确定性策略梯度

1：初始化批大小 M、时间步数 K 和经验回放缓冲区 D；初始化演员网络和评论家网络的在线网络（$\mu(s|\theta^\mu)$ 和 $Q(s, a|\theta^Q)$），其参数为 θ^μ 和 θ^Q；初始化目标网络（$\mu'(s|\theta^{\mu'})$ 和 $Q'(s, a|\theta^{Q'})$），其参数为 $\theta^{\mu'}$ 和 $\theta^{Q'}$；初始化队列状态和交通信号灯状态： $X_{n1}, X_{n2}, X_{n3}, X_{n4} = 0$；对所有 $n \in \{1, ..., N\}$，设置 $L_n = 0$； 2：对于步骤 $k = 1, ..., K$：3：选择动作 $a = \mu(s|\theta^\mu) + \mathcal{N}$，其中 $\mathcal{N}$ 为高斯噪声；4：执行动作 a，获得新状态 $s' = [X_{n1}, X_{n2}, X_{n3}, X_{n4}$；对所有 $n \in \{1, ..., N\}$，设置 $L_n]$，并从算法1获取奖励 r；5：将状态转移 (s, a, r, s') 存入内存；6：采样由 M 个状态转移组成的小批量 $\{(s_t^j, a_t^j, r_t^j, s_{t+1}^j)\}_j$；7：计算目标Q值： $y^j = r_t^j + \gamma Q'(s_{t+1}^j, \mu'(s_{t+1}^j|\theta^{\mu'})|\theta^{Q'})$；8：通过最小化损失函数更新评论家在线Q网络 Q ： $\nabla_{\theta^Q}\text{Loss}_t(\theta^Q) = \nabla_{\theta^Q}[\frac{1}{M}\sum_{j=1}^{M}(y^j - Q(s_t^j, a_t^j|\theta^Q))^2]$；9：通过链式法则，使用采样策略梯度更新演员网络的在线策略网络：$\mathbb{E}_{s_t}[\nabla_{\theta^\mu} Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t|\theta^\mu)}] = \mathbb{E}_{s_t}[\nabla_a Q(s, a|\theta^Q)|_{s=s_t, a=\mu(s_t)}\nabla_{\theta^\mu}\mu(s|\theta^\mu)|_{s=s_t}]$；10：使用 $\tau \ll 1$ 更新评论家的目标Q网络 Q' 和演员网络的目标策略网络 μ'：$\theta^{Q'} \leftarrow \tau\theta^Q + (1 - \tau)\theta^{Q'}$, $\theta^{\mu'} \leftarrow \tau\theta^\mu + (1 - \tau)\theta^{\mu'}$。

结合贝尔曼最优方程 [61] 和 $V_{\hat{\pi}} = \max \mathbb{E}[\sum_{t=0}^{T}\gamma^t|X(t)|^2]$，我们推导出以下方程

$$V_{\hat{\pi}} = \max\{r_t + \gamma \max \sum_{k=1}^{T-t}\gamma^{k-1} r_{t+k}\}, \tag{9}$$

其中 $r_t = -|X(t)|^2$。我们将 $Q(s, a)$ 定义为：在状态 $s = (X; L)$ 下采取动作 a 后，遵循最优策略所能获得的最大期望折扣奖励（在本文语境中即最小折扣交通拥堵成本）。$Q(s, a)$ 满足以下方程

$$\begin{aligned} Q(s, a) &= r(s, a) + \gamma \sum_{s' \in \mathcal{S}} p(s, s'; a) \max_{a' \in \mathcal{A}} Q(s', a') \\ &= r(s, a) + \gamma \mathbb{E}\left[\max_{a' \in \mathcal{A}} Q(s', a')\right], \end{aligned} \tag{10}$$

其中，$r(s, a) = -Z(X)$ 表示队列状态 X 下的奖励（即负的拥堵成本），而 $p(s, s'; a)$ 表示采取动作 a 时从状态 s 转移到状态 s' 的转移概率。

B. 单个道路交叉口的深度Q网络

在 DQN算法 [62],中，神经网络以状态和动作为输入，以动作价值为输出。与传统的 Q学习算法相比，深度Q网络使用神经网络作为函数近似器来估计动作价值函数。神经网络通过在每次迭代中调整其参数进行训练，以降低贝尔曼方程中的均方误差。

此外，DQN算法使用经验回放缓冲区来更新网络参数。它有两个优点：1）使随机梯度下降算法 [63] 得以实现；2）消除连续状态转移之间的相关性。

我们使用算法1模拟环境。第1行根据深度强化学习算法选择一个动作。第2行生成每种交通流的到达车辆数。第3行计算新的队列状态和交通信号灯状态。第4行计算奖励。最后，第5行更新所有状态。

算法2展示了单个道路交叉口的DQN细节。第1行为初始化。第 2 $\sim$ 7 行为训练过程中的 K步循环。第3行确定动作。第4行执行该动作，更新新的队列状态和交通信号灯状态，并获得奖励。第5行存储状态转移。第6行采样一个状态转移小批量，第7行更新网络。

C. 网格道路网络的可扩展深度确定性策略梯度

DDPG算法 [60] 采用神经网络，即评论家网络和演员网络，分别近似动作价值和动作策略。评论家网络以状态和动作为输入，以动作价值为输出。演员网络以状态为输入，以动作策略为输出。这种行动者—评论家方法有两项优势：1）它使用神经网络作为近似器，本质上将状态和动作空间压缩到小得多的潜在参数空间；2）可以使用梯度下降法更新网络权重，从而大幅加快收敛并缩短训练时间。因此，内存和计算资源都得以大量节省。DDPG算法在 AlphaGo Zero [64]和 Atari 游戏中展现出了卓越而强大的能力 [65]。

除了行动者—评论家方法外，DDPG算法主要使用三种技术来解决马尔可夫决策过程问题：经验回放缓冲区、软更新和探索噪声。经验回放缓冲区存储状态转移，其方式与 DQN算法中的描述相同。采用较低学习率进行软更新，以提高学习稳定性。在演员网络的目标策略中加入探索噪声，得到新的探索策略，从而有助于探索其他空间并获得更优策略。

首先，使用内存作为样本池对DDPG算法进行离线训练。在训练过程中，DDPG算法在线采样状态转移，并利用内存中收集的数据离线训练智能体，以进行进一步学习。网络使用Adam优化器进行 [66] 梯度下降更新。在测试过程中，DDPG算法以在线形式运行，不添加探索噪声。

传统DDPG算法具有连续动作空间，而我们模型中的交通控制动作是离散的，即 $A_n(t) \in \{0,1\}$。因此，我们采用离散化过程，将演员网络的连续输出转换为离散输出。在演员网络的输出层添加了一个经过修改的sigmoid函数作为激活函数：

$$\widehat{y} = \text{sigmoid}(\widehat{\alpha}\,\widehat{x}), \tag{11}$$

其中，$\widehat{x}$，$\widehat{y}$ 分别是最后一层的输入和输出，$\widehat{\alpha}$是用于增大 sigmoid 函数斜率的比例系数。通过将改进的 sigmoid 激活函数与逐节点二值化过程相结合，我们的离散DDPG算法可以大幅减少连续到离散转换所造成的误差。

为了在网格道路网络中收敛到一种较优策略，我们将所有交叉口的车辆数量和交通灯状态作为DDPG的输入。以交叉口n 为例。每个队列中的车辆数量为 $[X_{n1}, X_{n2}, X_{n3}, X_{n4}]$，灯状态为 L_n。观测元组 $S = [X_{n1}, X_{n2}, X_{n3}, X_{n4}, L_n]$ 是一个五元组，也是评论家在线Q网络和行动者的在线策略网络的输入。

算法3展示了网格道路网络中DDPG的详细过程。第1行进行初始化。在接下来的 for 循环中（第2行$\sim$ 6），我们训练策略。第3行根据行动者的在线策略网络和探索噪声选择一个动作。第4行执行该动作，并获取新的状态和奖励。第5行存储状态转移。第6行抽取一个状态转移小批量。第7行计算目标Q值，第8行更新评论家在线Q网络。第9行更新行动者的在线策略网络。第10行分别根据各自的在线网络更新评论家的目标Q网络 Q' 和行动者的目标策略网络 μ'。

D. “绿波”策略的最优性

在本小节中，我们证明了对称流体交通下网格道路网络中“绿波”控制策略的最优性。

为获得网格道路网络中的最优策略，我们首先求得单个交叉口的最优策略，然后将其扩展到网格道路网络，从而得到“绿波”策略。

在单交叉口场景中，最优策略是一种“阈值策略” [19]。只有当两个队列长度之差达到临界值时，交通信号灯才会切换。

在网格道路网络场景中，我们假设有一条大道（即一条主干道），并设有多条横向街道。我们将所有交叉口视为一个大型“交叉口”。“绿波”策略定义如下：仅当大道的队列长度与横向街道的队列长度之差达到临界值时，交通信号灯才会同时切换。对于大道上的车辆而言，它将依次遇到连续的绿灯，无需在任何交叉口处刹车。因此，这种控制策略被称为“绿波”。“绿波”策略在网格道路网络（即一条设有多条横向街道的大道）中实现了最优性。详细证明见附录。

V. 性能评估

本节首先介绍实验设置，然后展示单个道路交叉口中的结果，最后展示网格道路网络中的结果，尤其是“绿波”模式的出现。

A. 实验设置

在实验中，我们在单交叉口场景下的深度Q网络中采用一个2层全连接神经网络，其规模为400，并使用 tanh 激活函数。对于网格道路网络场景，我们在深度确定性策略梯度中采用一个4层全连接神经网络，演员网络和评论家网络的规模均为600。演员网络需要输出接近二值的动作值，因此在最后的输出层使用一个改进的 sigmoid 激活函数；而评论家网络的最后输出层不使用激活函数。

演员网络的输出被裁剪为二值0和1，分别表示交通灯状态的保持和切换。我们将仿真回合长度设为150步，以收集学习样本，并使用批大小为64和 γ =0.99训练演员网络和评论家网络。探索过程中采用方差为0.3的 Ornstein-Uhlenbeck 噪声 [67] 。

我们的算法使用深度学习框架 TensorFlow [68] 实现。仿真在一台服务器上运行，该服务器配备 Linux 操作系统、200 GB 内存、两颗 Intel(R) Xeon(R) Gold 5118 CPU（主频为 2.30 GHz）以及一块 Tesla V100-PCIE 图形处理器。

下面讨论交通环境设置。车辆到达率表示每个时隙中从道路网络外部进入道路网络、抵达各交叉口的车辆数量，设为一个随机值，上限分别为大道 C_a 、横向街道 C_c ；车辆通过率表示每个时隙内通过一个交叉口的车辆数量，大道和横向街道的车辆通过率分别固定为 16 和 4。到达/通行比是到达率与通过率之比。图4展示了不同策略下，单个交叉口的平均队列长度（随时间计算的所有道路上排队车辆总数的平均值）与到达/通行比之间的关系。

B. 验证单个道路交叉口中的最优控制

作为初步验证基准，我们首先考虑第 III-A 小节所述的单个道路交叉口场景。之所以考虑这一简单场景，是因为其状态-动作空间足够小，可以使用策略迭代等传统MDP方法计算最优策略。我们假设每个时间步两个交通流的到达车辆数由随机变量表示，这些随机变量相互独立，并服从参数为 $P = 0.25$ 的伯努利分布。

对于单个交叉口，最优策略是“阈值策略” [19]。并且，可以使用 MDP 工具箱 [69],中的策略迭代算法求得最优策略，该策略具有“阈值策略” [19]。各策略的比较如图3所示，包括 DQN算法、固定周期策略和最优策略得出的结果。DQN策略的结果与最优策略一致。由于单个道路交叉口的状态-动作空间非常小，使用 DQN 的训练时间极短，可以忽略不计。具体而言，它们与最优性能相当，并呈现出类似的阈值结构。对于一个密切相关的双队列动态优化问题（其中采用切换成本而非切换时间），这一结构特性已在 [19] 中得到证实。图4探讨了不同到达/通行比的影响。随着车辆到达/通过比增加，队列长度也随之增加。DQN学习到的策略比固定周期策略效果更好。图5研究了初始队列长度较大时的稳定化情况。到达/通行比为0.5。DQN学习到的策略能够稳定交通流。

C. 网格道路网络中的“绿波”

现在我们转向图 2 所示的网格道路网络。这是一个更具挑战性的场景，可用于检验我们的DDPG算法的可扩展性。此外，沿大道行驶的车辆流会产生更复杂的相互作用。

我们假设每个时间步大道东西方向的到达车辆数由随机变量表示，这些随机变量相互独立，并服从参数为 $P_1 = 0.5$ 的伯努利分布。 N 条横向街道中每条街道南北方向的到达车辆数由随机变量表示，这些随机变量相互独立，并服从参数为 $P_2 = 0.25$ 的伯努利分布。接下来，我们评估在 5×10 网格道路网络中的性能。

图6展示了随时间步变化的车辆数量。由于状态-动作空间的爆炸，传统MDP方法（如策略迭代）在计算上不可行，因此无法对我们的算法是否具有优势进行定量评估。因此，我们考察定性特征，以验证其智能行为并评估其性能优势。特别是，我们观察到了“绿波”模式的出现，尽管优化过程中并未明确规定此类结构特征。这种涌现的智能行为证实了我们的算法仅凭被动观测即可学习有利的结构属性的能力。

六、结论

我们探讨了深度Q网络（DQN）和深度确定性策略梯度（DDPG）算法在单个道路交叉口和网格道路网络中优化实时交通信号灯控制策略的潜力。作为初步基准，我们确定，在单个交叉口中，DQN算法获得的策略与传统MDP方法得到的最优策略相匹配。随后，我们评估了DDPG算法在网格道路网络中的可扩展性，并展示了名为“绿波”策略的高层智能行为的出现，证实了该算法学习理想结构特征的能力。我们

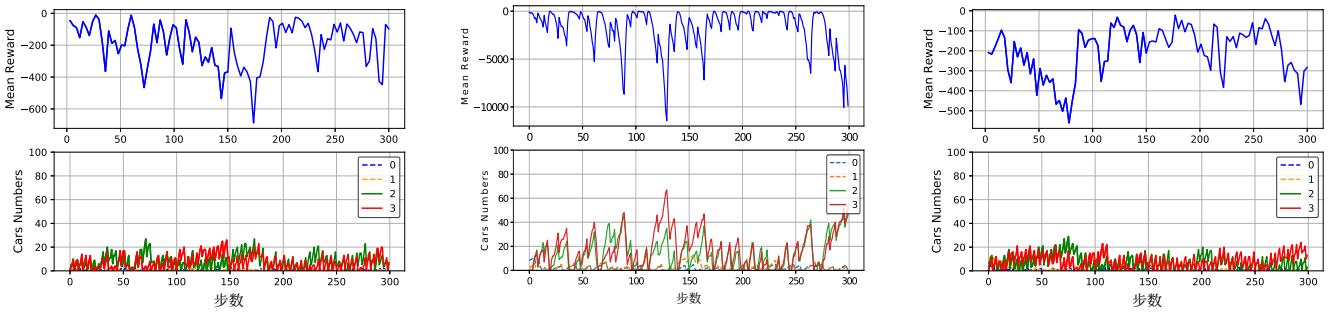

图3。单个交叉口中深度Q网络控制策略（左）、固定周期策略（中）和最优策略（右）的性能。图中展示了仿真中进行的300步测试，包括平均奖励和等待车辆数。更好的策略应能减少等待车辆数。DQN算法的表现与最优策略相当。

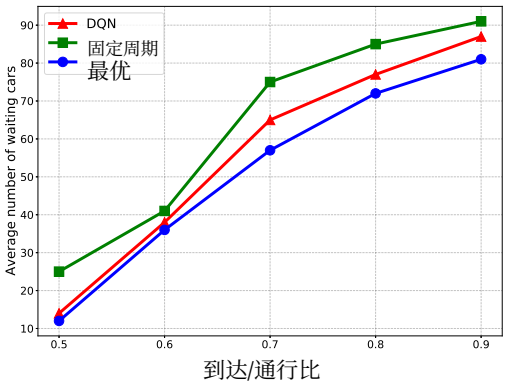

图4。单个交叉口在不同车辆到达/通过率下的比较。

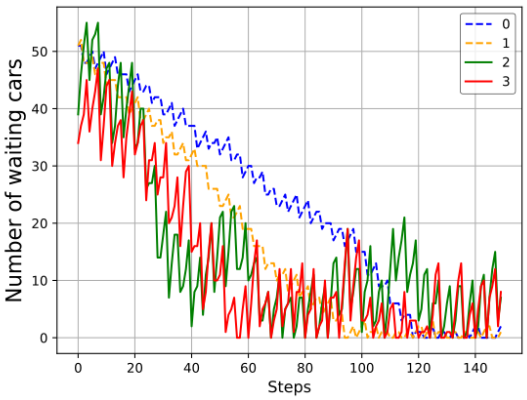

图5。初始队列长度较大时单个交叉口的稳定化。

还从理论上证明，在指定的交通流模型（具有多个横向街道的大道）下，“绿波”策略是最优控制策略。“绿波”模式表明，深度强化学习算法能够在网格道路网络中输出有利的解。

在未来的研究中，我们将考虑道路网络各区域之间车辆随机到达和离开的交通流，并计划研究局部性，分析如何利用新技术设计适用于大规模部署场景的分布式协调方案，例如多智能体深度强化学习算法、分层深度强化学习算法和边缘计算。

参考文献

[1] W. Hong、G. Tao、H. Wang 和 C. Wang，“采用多模型神经网络的自适应在线学习方案进行交通信号控制”，*IEEE* 神经网络与学习系统汇刊（*TNNLS*），2022年。[2] F.-X. Devailly、D. Larocque 和 L. Charlin，“Ig-rl：用于大规模交通信号控制的归纳式图强化学习”，*IEEE* 智能交通系统汇刊（*TITS*），2021年。[3] M. M. Vazifeh、P. Santi、G. Resta、S. Strogatz 和 C. Ratti，“解决按需城市出行中的最小车队问题”，《自然》，第557卷，第7706期，第534页，2018年。[4] M. Zhu、X.-Y. Liu、F. Tang、M. Qiu、R. Shen、W. Shu 和 M.-Y. Wu，“面向未来城市交通的公共车辆”，*IEEE* 智能交通系统汇刊（*TITS*），第17卷，第12期，第3344–3353页，2016年。

[5] “新研究强调经济与交通拥堵的关联。” https://mobility.tamu.edu/umr/media-information/press-release/，2021年。[6] T.-Q. Tang、Z.-Y. Yi 和 Q.-F. Lin，“信号灯对跟驰模型下燃料消耗和排放的影响”，爱思唯尔《物理学*A*辑：统计力学及其应用》，第 469 卷，第 200–205 页，2017年。[7] Y. Liu、S. Liu、Y. Wang、F. Lombardi 和 J. Han，“面向机器学习应用的随机计算神经网络综述”，*IEEE* 神经网络与学习系统汇刊（*TNNLS*），第 32 卷，第 7 期，第 2809–2824 页，2020年。[8] X. Liang、X. Du、G. Wang 和 Z. Han，“用于交通信号灯周期控制的深度强化学习网络”，*IEEE* 车辆技术汇刊（*TVT*），第 68 卷，第 2 期，第 1243–1253 页，2019年。[9] R. Mao、R. Cui 和 C. P. Chen，“具有强化学习信号反馈的宽度学习：理论与应用”，*IEEE* 神经网络与学习系统汇刊（*TNNLS*），2021年。[10] P. Koonce 和 L. Rodegerdts，“交通信号配时手册”，技术报告，美国联邦公路管理局，2008年。[11] R. P. Roess、E. S. Prassas 和 W. R. McShane，交通工程。培生/普伦蒂斯·霍尔，2004年。[12] J. D. Little、M. D. Kelson 和 N. H. Gartner，“Maxband：用于在主干道和三角形路网上设置信号灯的通用程序”，1981年。[13] S.-B. Cools、C. Gershenson 和 B. D’Hooghe，“自组织交通信号灯：一种现实仿真”，载于施普林格《自组织系统进展》，第 45–55 页，2013年。[14] P. Varaiya，“适用于任意信号控制交叉口网络的最大压力控制器”，载于施普林格《复杂交通系统中的动态网络建模进展》，第 27–66 页，2013年。[15] P. Lowrie，“SCATS——一种控制城市交通流的交通响应方法”，由澳大利亚悉尼道路与交通管理局出版的销售信息手册，1990年。[16] S. L¨ammer 和 D. Helbing，“交通信号灯和城市道路网络中车辆的自我控制

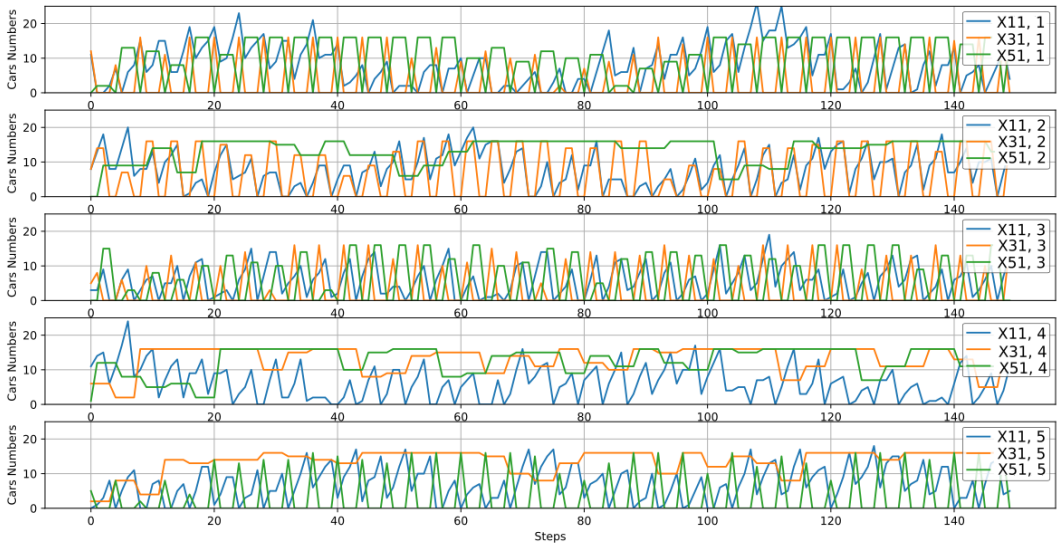

图6。适用于包含 5 × 10 个交叉口的网格道路网络的深度确定性策略梯度策略。

流”，《统计力学杂志：理论与实验》，第 2008 卷，第 04 期，第 P04019 页，2008年。[17] S. L¨ammer 和 D. Helbing，“现实饱和网络交通的自稳定分布式信号控制”，圣菲研究所，2010年。[18] I. M. Albatish 和 S. S. Abu-Naser，“使用基于规则的系统对智能交通信号灯系统进行建模与控制”，载于*IEEE* 新兴电子技术国际会议（*ICPET*），第 55–60 页，2019 年。[19] M. Hofri 和 K. W. Ross，“ 具有服务器设置时间的两个队列的最优控制及其分析”，《*SIAM* 计算杂志》，第 16 卷，第 2 期，第 399–420 页，1987年。[20] S. Onori、L. Serrao 和 G. Rizzoni，“动态规划”，载于《混合动力汽车》，第 41–49 页，施普林格，2016年。[21] M. L. Puterman，《马尔可夫决策过程：离散随机动态规划》。约翰·威立父子公司，2014年。[22] S. M. Ross，《随机动态规划导论》。学术出版社，2014年。[23] N. Kumar、S. S. Rahman 和 N. Dhakad，“模糊推理驱动的基于深度强化学习的智能交通系统交通信号灯控制”，*IEEE* 智能交通系统汇刊（*TITS*），2020年。[24] A. Haydari 和 Y. Yilmaz，“面向智能交通系统的深度强化学习：综述”，*IEEE* 智能交通系统汇刊（*TITS*），2020年。[25] V. Mnih、K. Kavukcuoglu、D. Silver、A. A. Rusu、J. Veness、M. G. Bellemare、A. Graves、M. Riedmiller、A. K. Fidjeland、G. Ostrovski、等，“通过深度强化学习实现人类水平控制”，《自然》，第 518 卷，第 7540 期，第 529 页，2015年。[26] T. Tan、F. Bao、Y. Deng、A. Jin、Q. Dai 和 J. Wang，“面向大规模交通网格信号控制的协同深度强化学习”，*IEEE* 控制论汇刊，第 50 卷，第 6 期，第 2687–2700 页，2019年。[27] W.-H. Lin 和 C. Wang，“用于交通信号控制的增强型 0-1 混合整数线性规划表述”，*IEEE* 智能交通系统汇刊（*TITS*），第 5 卷，第 4 期，第 238–245 页，2004年。[28] L. P. J. Rani、M. K. Kumar、K. Naresh 和 S. Vignesh，“使用红外线（IR）和物联网（IoT）的动态交通管理系统”，载于*IEEE* 科学、技术、工程与管理国际会议（*ICONSTEM*），第 353–357 页，2017年。[29] V. Kodire、S. Bhaskaran 和 H. Vishwas，“基于 GPS 和 Zigbee 的交通信号优先控制”，载于*IEEE* 创新计算技术国际会议（*ICICT*），第 2 卷，第 1–5 页，2016年。[30] S. Panichpapiboon 和 P. Leakkaw，“交通密度估计：一种移动感知方法”，《*IEEE* 通信杂志》，第 55 卷，第 12 期，第 126–131 页，2017年。[31] M. Zhu、X.-Y. Liu 和 X. Wang，“用于无人机辅助车联网的深度强化学习”，*IEEE* 智能交通系统汇刊（*TITS*），已投稿，第 15 页，2021年。[32] Z. Ma、M. Xiao、Y. Xiao、Z. Pang、H. V. Poor 和 B. Vucetic，“物联网的高可靠低时延无线通信：挑战、基础与赋能技术”，*IEEE*《物联网期刊》（*IOTJ*），第 6 卷，第 5 期，第 7946–7970 页，2019年。[33] H. Wu、Z. Zhang、C. Guan、K. Wolter 和 M. Xu，“利用分布式深度学习协作开展边缘与云计算，实现智慧城市物联网”，*IEEE*《物联网期刊》（*IOTJ*），第 7 卷，第 9 期，第 8099–8110 页，2020年。[34] H. Wang、P. Hu、P. Wang 和 H. Wang，“城市交通信号控制的遗传定时调度模型”，爱思唯尔《信息科学》，2021年。[35] A. M. Turky、M. S. Ahmad、M. Z. M. Yusoff，“遗传算法在交通信号灯与行人过街控制中的应用”，国际计算机科学与网络安全期刊，第 9 卷，第 2 期，第 88–96 页，2009年。[36] B. “Brian” Park、I. Yun 和 K. Ahn，“可持续交通信号控制的随机优化”，泰勒与弗朗西斯《国际可持续交通期刊》，第 3 卷，第 4 期，第 263–284 页，2009年。[37] M. E. M. ALI、A. DURDU、S. A. CELTEK 和 A. Yilmaz，“基于模糊逻辑并使用SUMO交通仿真器结合韦伯斯特公式和修正韦伯斯特公式的交通信号控制自适应方法”，*IEEE Access*，2021年。[38] R. Hawi、G. Okeyo 和 M. Kimwele，“利用模糊逻辑和无线传感器网络实现智能交通信号灯控制”，载于 *IEEE*计算会议，第 450–460 页，2017年。[39] P. W. Shaikh、M. El-Abd、M. Khanafer 和 K. Gao，“用于解决交通信号控制问题的群体智能与进化算法综述”，*IEEE* 智能交通系统汇刊（*TITS*），第 23 卷，第 1 期，第 48–63 页，2020年。[40] H. R. Maier、S. Razavi、Z. Kapelan、L. S. Matott、J. Kasprzyk 和 B. A. Tolson，“导论性概述：利用进化算法及其他元启发式算法进行优化”，爱思唯尔《环境建模与软件》，第 114 卷，第 195–213 页，2019年。[41] H. Wei、G. Zheng、H. Yao 和 Z. Li，“IntelliLight：一种用于智能交通信号灯控制的强化学习方法”，载于 *ACM SIGKDD*知识发现与数据挖掘国际会议，第 2496–2505 页，2018年。[42] T. Chu、J. Wang、L. Codec`a 和 Z. Li，“用于大规模交通信号控制的多智能体深度强化学习”，*IEEE* 智能交通系统汇刊（*TITS*），第 21 卷，1095，2019年3月，1086，第 期，第 – 页。

[43] P. Balaji、X. German 和 D. Srinivasan，“使用强化学习智能体进行城市交通信号控制”，IET智能交通系统，第 4 卷，第 3 期，第 177–188 页，2010年。[44] I. Arel、C. Liu、T. Urbanik 和 A. G. Kohls，“基于强化学习的用于网络交通信号控制的多智能体系统”，IET智能交通系统，第 4 卷，第 2 期，第 128–135 页，2010年。[45] P. Mannion、J. Duggan 和 E. Howley，“自适应交通信号控制强化学习算法的实验性综述”，施普林格《自主道路运输支持系统》，第 47–66 页，2016年。[46] B. Bakker、S. Whiteson、L. Kester 和 F. C. Groen，“通过多智能体强化学习系统进行交通信号灯控制”，载于施普林格《交互式协作信息系统》，第 475–510 页，2010年。[47] S. S. Mousavi、M. Schukat 和 E. Howley，“使用深度策略梯度和基于价值函数的强化学习进行交通信号灯控制”，IET智能交通系统，第 11 卷，第 7 期，第 417–423 页，2017年。[48] H. Wei、G. Zheng、H. Yao 和 Z. Li，“IntelliLight：一种用于智能交通信号灯控制的强化学习方法”，载于《第24届ACM SIGKDD知识发现与数据挖掘国际会议论文集》，第 2496–2505 页，美国计算机协会，2018年。[49] M. Aslani、M. S. Mesgari 和 M. Wiering，“现实交通网络中针对不同交通扰动事件的基于演员-评论家方法的自适应交通信号控制”，爱思唯尔《交通研究C辑：新兴技术》，第 85 卷，第 732–752 页，2017年。[50] M. Abdoos、N. Mozayani 和 A. L. Bazzan，“基于多智能体Q学习的非平稳环境交通信号灯控制”，载于 IEEE智能交通系统会议（ITSC），第 1580–1585 页，2011年。[51] E. Van der Pol 和 F. A. Oliehoek，“用于交通信号灯控制的协调式深度强化学习智能体”，《神经信息处理系统进展》（NeurIPS），第 1 卷，2016年。[52] X. Zang、H. Yao、G. Zheng、N. Xu、K. Xu 和 Z. Li，“Metalight：用于交通信号控制的基于价值的元强化学习”，载于《AAAI人工智能会议论文集》，第 34 卷，第 1153–1160 页，2020年。[53] C. Chen、H. Wei、N. Xu、G. Zheng、M. Yang、Y. Xiong、K. Xu 和 Z. Li，“迈向千灯：用于大规模交通信号控制的去中心化深度强化学习”，载于《AAAI人工智能会议论文集》（AAAI），第 34 卷，第 3414–3421 页，2020年。[54] T. Wu、P. Zhou、K. Liu、Y. Yuan 和 D. O. Wu，“车联网中用于城市交通灯控制的多智能体深度强化学习”，IEEE 车辆技术汇刊（TVT），第 PP 卷，第 99 期，第 1–1 页，2020年。[55] M. Xu、K. An、L. H. Vu、Z. Ye、J. Feng 和 E. Chen，“优化基于多智能体的城市交通信号控制系统”，Taylor & Francis《智能交通系统杂志》，第 23 卷，第 4 期，第 357–369 页，2019年。[56] W. Liu、J. Liu、J. Peng 和 Z. Zhu，“利用快速梯度下降函数近似实现车路协同网络的协作式多智能体交通信号控制系统”，载于 IEEE国际通信会议（ICC），第 2562–2567 页，2014年。[57] Y. Liu、L. Liu 和 W.-P. Chen，“使用分布式多智能体Q学习实现智能交通信号灯控制”，载于 IEEE智能交通系统国际会议（ITSC），第 1–8 页，2017年。[58] K. Zhang、Z. Yang 和 T. Bas¸ar，“多智能体强化学习：理论与算法精选综述”，Springer《强化学习与控制手册》，第 321–384 页，2021年。[59] Y. Yu、X. Si、C. Hu 和 J. Zhang，“循环神经网络综述：LSTM单元与网络架构”，《神经计算》，第 31 卷，第 7 期，第 1235–1270 页，2019年。[60] T. P. Lillicrap、J. J. Hunt、A. Pritzel、N. Heess、T. Erez、Y. Tassa、D. Silver 和 D. Wierstra，“通过深度强化学习实现连续控制”，载于学习表征国际会议（ICLR），2016年。[61] R. S. Sutton 和 A. G. Barto，《强化学习：导论》。MIT出版社，2018年。[62] V. Mnih、K. Kavukcuoglu、D. Silver、A. A. Rusu、J. Veness、M. G. Bellemare、A. Graves、M. Riedmiller、A. K. Fidjeland、G. Ostrovski 等，“通过深度强化学习实现人类水平控制”，《自然》，第 518 卷，第 7540 期，第 529 页，2015年。[63] A. Daniely，“随机梯度下降学习网络的共轭核类”，载于《神经信息处理系统进展》（NIPS），第 2422–2430 页，2017年。[64] D. Silver、J. Schrittwieser、K. Simonyan、I. Antonoglou、A. Huang、A. Guez、T. Hubert、L. Baker、M. Lai、A. Bolton 等，“无需人类知识掌握围棋”，《自然》，第 550 卷，第 7676 期，第 354 页，2017年。

[65] V. Mnih、K. Kavukcuoglu、D. Silver、A. Graves、I. Antonoglou、D. Wier- stra 和 M. Riedmiller，“使用深度强化学习玩 Atari 游戏”，*arXiv* 预印本 *arXiv:1312.5602*，2013年。[66] D. P. Kingma 和 J. Ba，“Adam：一种随机优化方法”，*arXiv* 预印本 *arXiv:1412.6980*，2014 年。[67] W. Horsthemke 和 R. Lefever，“ 外部宽带马尔可夫噪声的摄动展开：由 Ornstein-Uhlenbeck 噪声诱导的状态转移的应用”，施普林格 *Zeitschrift f¨ur Physik B Condensed Matter*，第 40 卷，第 3 期，第 241–247 页，1980 年。[68] M. Abadi、P. Barham、J. Chen、Z. Chen、A. Davis、J. Dean、M. Devin、S. Ghemawat、G. Irving、M. Isard、等，“TensorFlow：一个用于大规模机器学习的系统”，载于 *USENIX* 操作系统设计与实现研讨会 *(OSDI)*，第 265–283 页，2016年。[69] “Python 马尔可夫决策过程（MDP）工具箱（2022年）。” https://pymdptoolbox.readthedocs.io/en/latest/api/mdptoolbox.html。



Xiao-Yang Liu 于2018年获得美国纽约哥伦比亚大学电气工程理学硕士学位，目前是美国纽约哥伦比亚大学电气工程系的哲学博士候选人。

他的研究兴趣包括张量与张量网络、高性能张量计算、高性能张量计算、深度学习、非凸优化、大数据分析以及物联网应用。



Ming Zhu 在中国上海交通大学获得计算机科学与工程哲学博士学位。现任中国科学院深圳先进技术研究院助理教授，工作地点在中国深圳。

他的研究兴趣包括自动驾驶、智能交通系统、大数据、深度学习和张量网络。



Sem Borst Sem Borst于1990年获得特文特大学应用数学理学硕士学位，并于1994年获得蒂尔堡大学哲学博士学位。1994年，他曾在英格兰剑桥大学统计实验室担任访问学者。1995年，Sem加入美国默里希尔贝尔实验室网络与系统数学研究部。除了在埃因霍温理工大学的职位外，他还与贝尔实验室保持（兼职）隶属关系。Sem已在同行评审期刊和会议论文集中发表190多篇论文，并

他的H指数为39。他曾担任或现任多家期刊的编委会成员，例如 ACM计算系统建模与性能评估汇刊、IEEE/ACM《网络汇刊》、《运筹学数学方法》和《排队系统》，并曾担任多个会议的程序委员会委员。



Anwar Walid Anwar Walid（F'09）在美国纽约州纽约市的纽约大学获得电气与计算机工程理学学士和理学硕士学位，并在美国纽约州纽约市的哥伦比亚大学获得哲学博士学位。

他是网络智能总监、分布式系统研究员，也是诺基亚贝尔实验室美国新泽西州默里希尔分部的杰出研究人员；在那里，他还曾担任系统研究数学系主任和大学研究合作总监。

他是哥伦比亚大学电气工程系的兼职教授。他在网络技术和计算的各个方面拥有20多项美国及国际授权专利。他目前的研究兴趣包括分布式系统的控制与优化，以及学习模型与算法在物联网、数字健康、智能交通、云计算和软件定义网络中的应用。Walid博士曾获得IEEE和美国计算机协会颁发的奖项，包括2017年IEEE通信学会威廉·R·贝内特奖以及美国计算机协会SIGMETRICS/国际信息处理联合会（IFIP）性能最佳论文奖。他曾担任IEEE/ACM《云计算汇刊》、IEEE《网络》杂志以及IEEE/ACM《网络汇刊》的副编辑。他曾担任IEEE INFOCOM技术程序委员会主席、2018年IEEE/ACM互联健康会议（CHASE）大会主席，以及IEEE《物联网期刊》"面向物联网的人工智能赋能认知通信与网络"专刊的客座编辑。他是IFIP第7.3工作组和Tau Beta Pi工程荣誉学会的当选成员。

附录

我们首先描述交通模型。然后讨论稳定性条件和队列长度的演变。最后，证明"绿波"控制策略的最优性。

A. 模型描述

考虑一个具有 N 个节点的网格道路网络拓扑，该网络代表一条设有 N 条横向街道的大道（即一条主干道）。大道交通流以恒定速率 λ_0 作为确定性"流体流"到达节点 1，并必须经过所有节点，最后通过节点 N 离开网络。此外，横向交通流以恒定速率 λ_n 作为确定性流体流到达各节点 n，并在经过该节点后离开网络，$n \in \{1,...,N\}$。为保证稳定性，我们假设对所有 $n \in \{1,...,N\}$，$\lambda_0 + \lambda_n < 1$。

只有当交通信号灯为绿灯时，大道交通流才能通过节点 n：若在节点 n处排队，则以单位速率通过，$n \in \{1,...,N\}$；否则，以其上游节点 $n-1$的流出率通过（当 $n > 1$时），或以外部到达率 λ_0通过（当 $n = 1$时）。同样，只有当交通信号灯为红灯时，横向交通流才能通过节点 n：若有排队，则以单位速率通过；否则，以到达率 λ_n通过，$n \in \{1,...,N\}$。

分别用 Y_n 和 O_n 表示节点 $n \in \{1,...,N\}$ 处黄灯和橙灯时段的（固定）持续时间。节点 $n \in \{1,...,N\}$ 处第 k 个绿灯和红灯时段的长度可控，分别记为 $G_n^{(k)}$ 和 $R_n^{(k)}$。为简洁起见，用 $U_n^{(k)} = G_n^{(k)} + R_n^{(k)} + Y_n + O_n$ 表示节点 $n \in \{1,...,N\}$ 处第 k 个控制周期的长度。此外，定义

$$G_n = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} G_n^{(k)},\ R_n = \lim_{K\to\infty} \frac{1}{K}\sum_{k=1}^{K} R_n^{(k)}, \quad (12)$$

为节点 n 处绿灯和红灯时段的长期平均持续时间（假设这些极限存在）。用 U_n 表示节点 n 处的长期平均控制周期长度，

$$U_n = G_n + R_n + Y_n + O_n,\ n \in \{1,...,N\}. \quad (13)$$

B. 稳定性条件

为确保队列的稳定性，必须满足

$$\frac{G_n}{U_n} \geq \lambda_0,\quad \frac{R_n}{U_n} \geq \lambda_n,\quad n \in \{1,...,N\}, \quad (14)$$

因为主干道交通流以速率 λ_0 持续到达，并且只能在该节点的绿灯时段以单位速率通过节点 n；同样，节点 n 处的横向交通流以速率 λ_n 持续到达，并且只能在该节点的红灯时段以单位速率通过。

在附录的其余部分中，我们假设（14）中的不等式成立。根据（13），（14）中的两个不等式可等价地表示为

$$\begin{aligned} G_n &\geq \frac{\lambda_0}{1-\lambda_0}\left(R_n + Y_n + O_n\right),\ n \in \{1,...,N\},\\ R_n &\geq \frac{\lambda_n}{1-\lambda_n}\left(G_n + Y_n + O_n\right),\ n \in \{1,...,N\}, \end{aligned} \quad (15)$$

或者表示为以下形式

$$\begin{aligned} R_n &\leq \frac{1-\lambda_0}{\lambda_0} G_n - Y_n - O_n,\ n \in \{1,...,N\},\\ G_n &\leq \frac{1-\lambda_n}{\lambda_n} R_n - Y_n - O_n,\ n \in \{1,...,N\}. \end{aligned} \quad (16)$$

考虑到

$$1 - \frac{G_n}{U_n} - \frac{R_n}{U_n} = \frac{Y_n + O_n}{U_n} \overset{(14)}{\leq} 1 - \lambda_0 - \lambda_n, \quad (17)$$

因此，

$$U_n \geq \frac{Y_n + O_n}{1 - \lambda_0 - \lambda_n}. \quad (18)$$

此外，由（14）和（18）可得

$$G_n \geq \frac{\lambda_0}{1-\lambda_0-\lambda_n}(Y_n + O_n), \quad (19)$$

$$R_n \geq \frac{\lambda_n}{1-\lambda_0-\lambda_n}(Y_n + O_n). \quad (20)$$

注意，（13）和（14）结合可得

$$\lambda_0 + \lambda_n \leq 1 - \frac{Y_n + O_n}{U_n} < 1, \quad (21)$$

如上所假设，下面的引理还给出一些有用的不等式，其证明见附录。

引理1。对于 $n \in \{1,...,N\}$，

$$\frac{(G_n + Y_n + O_n)^2}{U_n} \geq \frac{\lambda_0(p_n+1)^2(Y_n + O_n)}{p_n}, \quad (22)$$

$$\frac{(R_n + Y_n + O_n)^2}{U_n} \geq \frac{\lambda_n(q_n+1)^2(Y_n + O_n)}{q_n}, \quad (23)$$

$$p_n = \max\{1, \frac{\lambda_0}{1-\lambda_0-\lambda_n}\},\quad q_n = \max\{1, \frac{\lambda_n}{1-\lambda_0-\lambda_n}\}.$$

上述不等式均为严格不等式，除非分别满足 $G_n = p_n($ $Y_n + O_n)$和 $R_n = q_n(Y_n + O_n)$。

证明。我们仅给出 (22) 的证明，因为 (23) 的证明类似。我们有

$$\frac{(G_n+Y_n+O_n)^2}{U_n} \overset{(14)}{\geq} \frac{\lambda_0(G_n+Y_n+O_n)^2}{G_n}.$$

求导可知，对于 $G_n \geq Y_n+O_n$，右侧随G_n 增大而增大；对于 $G_n < Y_n+O_n$，右侧随 G_n 增大而减小。因此，结合 (19) 中 G_n的下界，当 $p_n = \max\{1, \frac{\lambda_0}{1-\lambda_0-\lambda_n}\}$ 时，右侧的最小值在 $G_n = p_n(Y_n+O_n)$ 处取得，且等于

$$\frac{\lambda_0(p_n+1)^2(Y_n+O_n)}{p_n},$$

由此得到下界 (22)。还可知，除非 $G_n = p_n(Y_n+O_n)$，否则该下界是严格的。 □

C. 队列演化

记$\Phi_n(b,t)$ 和 $\Psi_n(b,t)$ 分别表示节点 n 处大道和横向交通流在 t 时刻的队列长度，其中交通流始于时刻 b。令 $\Phi_n(b,t)$ 和 $\Psi_n(b,t)$ 分别为 $\Phi_n(b,t)$ 和 $\Psi_n(b,t)$ 的下界。若交通流始于时刻 $b=0$，则 $\Phi_n(b,t)$、$\Psi_n(b,t)$、$\Phi_n(b,t)$ 和 $\Psi_n(b,t)$ 可分别简写为 $\Phi_n(t)$、$\Psi_n(t)$、$\Phi_n(t)$ 和 $\Psi_n(t)$。记 $G_n(b,t)$ 和 $R_n(b,t)$分别为节点 $n \in \{1,...,N\}$在时间区间 $(b,t]$ 内的非绿灯时间和非红灯时间。现在注意到，在节点 1 的非绿灯时段内，$\Phi_1(b,t)$ 以速率 λ_0 增加，而在绿灯时段内以速率 $1-\lambda_0$ 减少。因此，

$$\begin{aligned}\widehat{\Phi}_1(b,t) &\geq \lambda_0\overline{G}_1(b,t)-(1-\lambda_0)(t-b-\overline{G}_1(b,t))\\ &= \overline{G}_1(b,t)-(1-\lambda_0)(t-b),\end{aligned} \tag{24}$$

对所有 $b \leq t$，得到

$$\widehat{\Phi}_1(b,t) \geq \sup_{s\leq t}\{\overline{G}_1(b,t)-(1-\lambda_0)(t-b)\}, \tag{25}$$

并假设 $\Phi_1(0)=0$，实际上我们有

$$\Phi_1(t) = \sup_{s\leq t}\{\overline{G}_1(b,t)-(1-\lambda_0)(t-b)\}. \tag{26}$$

注释1. 不失一般性，我们假设每个控制周期都以绿灯时段结束，因此以黄灯时段开始。令 $B^{(k)}$ 为节点 *1* 处第 k个控制周期的开始时间。可以证明

$$\begin{aligned}\Phi_1(t) &= \sup_{s\leq t}\{\overline{G}_1(b,t)-(1-\lambda_0)(t-b)\}\\ &= \max\{\sup_{l\in\{1,...,\kappa(t)\}}\{\overline{G}_1(B^{(l)},t)-(1-\lambda_0)(t-B^{(l)})\},0\},\end{aligned} \tag{27}$$

其中 $\kappa(t)$ 是截至时间 t的最大周期数。

首先，如果 $b \in [B^{(l-1)}+G_1^{(l-1)}, B^{(l)}]$，则 $G_1(b,t) = \overline{G}_1(B^{(l)},t)$，且 $G_1(b,t)-(1-\lambda_0)(t-b) \leq G_1(B^{(l)},t)-(1-\lambda_0)(t-B^{(l)})$。其次，如果 $b \in [B^{(l)}, B^{(l)}+G_1^{(l)}]$，则 $G_1(b,t) = \overline{G}_1(B^{(l)},t)-(b-B^{(l)})$，其中 $b-B^{(l)} \geq 0$，且 $G_1(b,t)-(1-\lambda_0)(t-b) \leq G_1(B^{(l)},t)-(1-\lambda_0)(t-B^{(l)})$。最后，将上述两种情况合并为：对于所有 $b \in [B^{(l-1)}+G_1^{(l-1)}, B^{(l)}+G_1^{(l)}]$，$G_1(b,t)-(1-\lambda_0)(t-b) \leq G_1(B^{(l)},t)-(1-\lambda_0)(t-B^{(l)})$。我们看到，$\overline{G}_1(b,t)-(1-\lambda_0)(t-b) \leq G_1(B^{(l)},t)-(1-\lambda_0)(t-B^{(l)})$ 对每个周期 $l \in \{1,...,\kappa(t)\}$ 都始终成立。

同样，在节点 1 处，$R_1(b,t)$ 在非红灯时段以速率 λ_n 增长，在红灯时段以速率 $1-\lambda_n$ 减小。因此，

$$\begin{aligned}\widehat{\Psi}_n(b,t) &\geq \lambda_n\overline{R}_n(b,t)-(1-\lambda_n)(t-b-\overline{R}_n(b,t))\\ &= \overline{R}_n(b,t)-(1-\lambda_n)(t-b),\end{aligned} \tag{28}$$

对所有 $b \leq t$，得到

$$\widehat{\Psi}_n(t) \geq \sup_{s\leq t}\{\overline{R}_n(b,t)-(1-\lambda_n)(t-b)\}, \tag{29}$$

并假设 $\Psi_n(0)=0$，实际上我们有

$$\Psi_n(t) = \sup_{s\leq t}\{\overline{R}_n(b,t)-(1-\lambda_n)(t-b)\}. \tag{30}$$

$\Phi_n(t)$ 的动态变化与之相似，但也有所不同，因为节点 n 处主干道交通流的流入率并非固定，而是等于上游节点 $n-1$的流出率。用 $\eta_n(t)$ 表示时刻 t节点 n 处主干道交通流的流出率，并注意到，在非绿灯时段，$\eta_n(t)=0$；在绿灯时段，当 $\Phi_n(t)>0$时，$\eta_n(t)=1$；在绿灯时段，$\eta_n(t)=\eta_{n-1}(t)$；并且在所有时段，当 $\Phi_n(t)=0$ 时，$\eta_n(t)=0$。假设 $\Phi_n(0)=0$，我们有

$$\Phi_n(t) = \sup_{s\leq t}\{\int_{x=0}^{t}\eta_{n-1}(x)dx-(t-b-\overline{G}_n(b,t))\}. \tag{31}$$

定义

$$\Phi_n = \lim\sup_{T\to\infty}\frac{1}{T}\int_{t=0}^{T}\Phi_n(t)dt,$$
$$\Psi_n = \lim\sup_{T\to\infty}\frac{1}{T}\int_{t=0}^{T}\Psi_n(t)dt,$$

分别表示节点 n处主干道和横向交通流的长期平均队列长度。$\Phi_1(t)$ 的下界由引理 2 得出。

引理 2。

$$\Phi_1(t) \geq \sum_{k=1}^{\infty}\max\{\overline{G}_1^{(k)}(t)-(1-\lambda_0)U^{(k)},0\}, \tag{32}$$

进一步推导为

$$\Phi_1(t) \geq \sum_{k=1}^{\infty}\max\{\overline{G}_1^{(k)}(t)-(1-\lambda_0)|t-B^{(k)}|,0\}. \tag{33}$$

证明。不失一般性，我们假设每个控制周期都以绿灯时段结束（因此以黄灯时段开始）。为简洁起见，用 $G_1^{(k)} = R_1^{(k)}+Y_1+O_1$表示节点 1 在第 k个控制周期内的非绿灯时间。用 $B^{(k)}$ 表示节点 1 第 k个控制周期的开始时间。定义

$$\overline{G}_1^{(k)}(t) = \max\{\min\{\overline{G}_1^{(k)}, t-B^{(k)}\},0\},$$

即节点 1 在第 k 个控制周期内，截至时间 t已经经过的非绿灯时间量，或者等价地，

$$\overline{G}_1^{(k)}(t) = \begin{cases}0, & \text{if } t \leq B^{(k)},\\ t-B^{(k)}, & \text{if } B^{(k)} \leq t \leq B^{(k)}+\overline{G}_1^{(k)},\\ \overline{G}_1^{(k)}, & \text{if } t \geq B^{(k)}+\overline{G}_1^{(k)}.\end{cases} \tag{34}$$

定义 $G_1^{(k)}(b,t) = G_1^{(k)}(t) - G_1^{(k)}(b)$ 为节点 1 在时间区间 $(b,t]$ 内第 k 个控制周期所对应的非绿灯时间量，并注意 $G_1(b,t)\sum_{k=1}^{\infty} G_1^{(k)}\ b\ t$ (,) 是节点 1 在时间区间 $(b,t]$ 内的非绿灯时间总量。令 $U^{(k)}$ 为第 k 个周期的持续时间。

$$\begin{aligned}\Phi_1(t) &\overset{(27)}{=} \sup_{l\in\{1,\dots,\kappa(t)\}}\{\overline{G}_1(B^{(l)},t)-(1-\lambda_0)(t-B^{(l)})\}\\ &\geq \overline{G}_1^{(1)}(B^{(1)},t)-(1-\lambda_0)(t-B^{(1)})\\ &= \sum_{k=1}^{\kappa(t)}[\overline{G}_1^{(k)}(t)-(1-\lambda_0)U^{(k)}] \qquad (35)\\ &\geq \sum_{k=1}^{\kappa(t)}[\overline{G}_1^{(k)}(t)-(1-\lambda_0)(t-B^{(k)})]. \qquad (36)\end{aligned}$$

当 $k \geq \kappa(t)+1, G_1^{(k)}(t) = 0$，且 $G_1^{(k)}(t)-(1-\lambda_0)U^{(k)} < 0$ 时。因此，结合式 (35)，可得式 (32)。当 $k = \kappa(t)+1$，$t - B^{(k)} = 0$。当 $k \geq \kappa(t)+2$，$t - B^{(k)} < 0$，且 $|t - B^{(k)}| > 0$ $\sum_{k=1}^{\kappa(t)}[G_1^{(k)} t - 1-\lambda_0\ t -$。因 $B^{(k)}\ ] = \sum_{k=1}^{\infty}\ \{G_1^{(k)}\ t\ -\ 1-\lambda_0\ |t-B^{(k)}|\ 0\}$ 此，( ) ( )( ) max ( ) ( )，。结合式 (36)，可得式 (33)。

现在注意，当 $B^{(k)} \leq t \leq B^{(k)} + \frac{1}{1-\lambda_0}G_1^{(k)}$ 时，我们有

$$\begin{aligned}&\max\{\overline{G}_1^{(k)}(B^{(k)},t)-(1-\lambda_0)|t-B^{(k)}|,0\}\\ &= \overline{G}_1^{(k)}(B^{(k)},t)-(1-\lambda_0)(t-B^{(k)})\\ &= \begin{cases}\lambda_0(t-B^{(k)})-\overline{G}_1^{(k)}(B^{(k)}), & \text{if } B^{(k)} \leq t \leq B^{(k)}+\overline{G}_1^{(k)},\\ \overline{G}_1^{(k)}-(1-\lambda_0)(t-B^{(k)}) & \\ \quad -\overline{G}_1^{(k)}(B^{(k)}), & \text{if } B^{(k)}+\overline{G}_1^{(k)} \leq t\\ & \leq B^{(k)}+\frac{1}{1-\lambda_0}\overline{G}_1^{(k)}.\end{cases}\end{aligned}$$

□

命题1。对于 $n \in \{1,...,N\}$，我们有

$$\begin{aligned}\Phi_1 &\geq \frac{\lambda_0(R_1+Y_1+O_1)^2}{2(1-\lambda_0)U_1},\\ \Psi_n &\geq \frac{\lambda_n(G_n+Y_n+O_n)^2}{2(1-\lambda_n)U_n}.\end{aligned} \qquad (37)$$

证明。定义

$$E^{(k)} = \overline{G}_1^{(k)} - (1-\lambda_0)U^{(k)},$$

and

$$F^{(k)} = \sup_{l\in\{1,\dots,k\}}\sum_{m=l}^{k} E^{(m)},$$

并考虑满足 $F^{(k)} = 0$ 的 k，即，对于所有 $l \in \{1,...,k\}$，均有 $\sum_{m=l}^{k} E^{(m)} \leq 0$，这意味着

$$\begin{aligned}\overline{G}_1^{(l)} - (1-\lambda_0)\sum_{m=l}^{k}U^{(m)} &\leq \sum_{m=l}^{k}\overline{G}_1^{(m)} - (1-\lambda_0)\sum_{m=l}^{k}U^{(m)}\\ &= \sum_{m=l}^{k}E^{(m)} \leq 0,\end{aligned} \qquad (38)$$

即，

$$\overline{G}_1^{(l)} \leq (1-\lambda_0)\sum_{m=l}^{k}U^{(m)},$$

因此

$$B^{(l)} + \frac{1}{1-\lambda_0}\overline{G}_1^{(l)} \leq B^{(l)} + \sum_{m=l}^{k}U^{(m)} = B^{(k+1)}.$$

现在考虑

$$\begin{aligned}&\int_{t=0}^{B^{(k+1)}}\Phi_1(t)dt\\ &\overset{(33)}{\geq} \int_{t=0}^{B^{(k+1)}}\sum_{l=1}^{\kappa(t)}\max\{\overline{G}_1^{(l)}(t)-(1-\lambda_0)|t-B^{(l)}|,0\}dt\\ &= \sum_{l=1}^{\kappa(t)}\int_{t=0}^{B^{(k+1)}}\max\{\overline{G}_1^{(l)}(t)-(1-\lambda_0)|t-B^{(l)}|,0\}dt.\end{aligned} \qquad (39)$$

注意

$$\begin{aligned}&\max\{\overline{G}_1^{(l)}(t)-(1-\lambda_0)|t-B^{(l)}|,0\} =\\ &\begin{cases}0, & \text{if } t \leq B^{(l)},\\ \lambda_0(t-B^{(l)}), & \text{if } B^{(l)} \leq t \leq B^{(l)}+\overline{G}_1^{(l)},\\ \overline{G}_1^{(l)}-(1-\lambda_0)(t-B^{(l)}), & \text{if } B^{(l)}+\overline{G}_1^{(l)} \leq t\\ & \leq B^{(l)}+\frac{1}{1-\lambda_0}\overline{G}_1^{(l)},\\ 0, & \text{if } t \geq B^{(l)}+\frac{1}{1-\lambda_0}\overline{G}_1^{(l)}.\end{cases}\end{aligned} \qquad (40)$$

Thus

$$\begin{aligned}&\int_{t=0}^{B^{(k+1)}}\max\{\overline{G}_1^{(l)}(t)-(1-\lambda_0)t,0\}\\ &= \int_{t=B^{(l)}}^{B^{(l)}+\overline{G}_1^{(l)}}\lambda_0(t-B^{(l)})dt\\ &\quad + \int_{t=B^{(l)}+\overline{G}_1^{(l)}}^{B^{(l)}+\frac{1}{1-\lambda_0}\overline{G}_1^{(l)}}[\overline{G}_1^{(l)}-(1-\lambda_0)(t-B^{(l)})]dt\\ &= \int_{x=0}^{\overline{G}_1^{(l)}}\lambda_0 x dx + \int_{x=\overline{G}_1^{(l)}}^{\frac{1}{1-\lambda_0}\overline{G}_1^{(l)}}[\overline{G}_1^{(l)}-(1-\lambda_0)x]dx\\ &= \left[\frac{1}{2}\lambda_0 x^2\right]_0^{\overline{G}_1^{(l)}} + \frac{\lambda_0\left(\overline{G}_1^{(l)}\right)^2}{1-\lambda_0} - \left[\frac{1}{2}(1-\lambda_0)x^2\right]_{\overline{G}_1^{(l)}}^{\frac{1}{1-\lambda_0}\overline{G}_1^{(l)}}\\ &= \frac{1}{2}\lambda_0\left(\overline{G}_1^{(l)}\right)^2 + \frac{\lambda_0\left(\overline{G}_1^{(l)}\right)^2}{1-\lambda_0} - \frac{1}{2}\frac{\left(\overline{G}_1^{(l)}\right)^2}{1-\lambda_0}\\ &\quad + \frac{1}{2}(1-\lambda_0)\left(\overline{G}_1^{(l)}\right)^2\\ &= \frac{\lambda_0\left(\overline{G}_1^{(l)}\right)^2}{2(1-\lambda_0)}.\end{aligned} \qquad (41)$$

我们得到

$$\int_{t=0}^{B^{(k+1)}}\Phi_1(t)dt \geq \sum_{l=1}^{\kappa(t)}\frac{\lambda_0\left(\overline{G}_1^{(l)}\right)^2}{2(1-\lambda_0)} = \frac{\lambda_0}{2(1-\lambda_0)}\sum_{l=1}^{\kappa(t)}\left(\overline{G}_1^{(l)}\right)^2.$$

现在注意到

$$\lim_{K\to\infty}\frac{1}{K}\sum_{k=1}^{K}E^{(k)} = R_1+Y_1+O_1-(1-\lambda_0)U_1 = -G_1+\lambda_0U_1<0, \tag{42}$$

这意味着 k 中使 $F^{(k)}=0$成立的子序列是无限的，因此

$$\begin{aligned}\Phi_1 &= \limsup_{T\to\infty}\frac{1}{T}\int_{t=0}^{T}\Phi_1(t)dt\\ &\geq \limsup_{k\to\infty}\frac{1}{B^{(k+1)}}\int_{t=0}^{B^{(k+1)}}\Phi_1(t)dt\\ &\geq \limsup_{k\to\infty}\frac{1}{B^{(k+1)}}\frac{\lambda_0}{2(1-\lambda_0)}\sum_{l=1}^{k}\left(\overline{G}_1^{(l)}\right)^2\\ &= \limsup_{k\to\infty}\frac{1}{\sum_{l=1}^{k}U^{(l)}}\frac{\lambda_0}{2(1-\lambda_0)}\sum_{l=1}^{k}\left(\overline{G}_1^{(l)}\right)^2\\ &= \frac{\lambda_0}{2(1-\lambda_0)}\limsup_{k\to\infty}\frac{1}{\frac{1}{k}\sum_{l=1}^{k}U^{(l)}}\frac{1}{k}\sum_{l=1}^{k}\left(\overline{G}_1^{(l)}\right)^2\\ &= \frac{\lambda_0}{2(1-\lambda_0)}\frac{(R_1+Y_1+O_1)^2}{U_1},\end{aligned} \tag{43}$$

如所述。 □

结合引理1和命题1，我们得到下界

$$\begin{aligned}\Phi_1 &\geq \frac{\lambda_0}{2(1-\lambda_0)}\frac{\lambda_0(p_1+1)^2(Y_1+O_1)}{p_1}\\ &= \frac{\lambda_0^2(p_1+1)^2(Y_1+O_1)}{2p_1(1-\lambda_0)},\end{aligned}$$

$$\begin{aligned}\Psi_n &\geq \frac{\lambda_n}{2(1-\lambda_n)}\frac{\lambda_n(q_n+1)^2(Y_n+O_n)}{q_n}\\ &= \frac{\lambda_n^2(q_n+1)^2(Y_n+O_n)}{2q_n(1-\lambda_n)}, \qquad n\in\{1,...,N\},\end{aligned}$$

并得出结论：除非分别满足 $G_1=p_1(Y_1+O_1)$ 和 $R_n=q_n(Y_n+O_n)$，否则这两个不等式均为严格不等式。

此后，我们作一个温和的假设：所有节点的黄灯时段和橙灯时段的固定持续时间相同，即对所有 $n\in\{1,...,N\}$，$Y_n\equiv Y$和 $O_n\equiv O$。上述下界可简化为

$$\Phi_1 \geq \frac{\lambda_0^2(p_1+1)^2(Y+O)}{2p_1(1-\lambda_0)}, \tag{44}$$

$$\Psi_n \geq \frac{\lambda_n^2(q_n+1)^2(Y+O)}{2q_n(1-\lambda_n)}, \tag{45}$$

分别在 $G_1=p_1(Y_1+O_1)$ 和 $R_n=q_n(Y_n+O_n)$ 的情况下不等式取严格形式。

D. “绿波”控制策略的最优性

现在考虑一种绿波控制策略，它使所有节点的状态保持同步，也就是说，同时采用绿灯、橙色、红灯和黄灯时段。将$\lambda_{\max}=\max_{n\in\{1,...,N\}}\lambda_n$ 记为所有横向街道的最大到达率。我们选取一个统一的绿灯时段长度

$$G_n\equiv G(\delta)=\frac{\lambda_0(1+\delta)(Y+O)}{1-\lambda_0-\lambda_{\max}}, \tag{46}$$

以及相同长度的红灯时段

$$R_n\equiv R(\delta)=\frac{\lambda_{\max}(1+\delta)(Y+O)}{1-\lambda_0-\lambda_{\max}}, \tag{47}$$

其中，$\delta\geq 0$ 是基于下界 (19) 和 (20) 的一个较小可扩展因子。由此，(46) 和 (47) 得到一个共同的控制周期长度

$$U_n\equiv U(\delta)=\frac{(1+\delta(\lambda_0+\lambda_{\max}))(Y+O)}{1-\lambda_0-\lambda_{\max}}. \tag{48}$$

接下来的命题对于确立绿波控制策略的最优性起着关键作用。

命题2. 采用绿波控制策略时的长期平均队列长度等于

$$\begin{aligned}\Phi_1(\delta) &= \frac{\lambda_0(R(\delta)+Y+O)^2}{2(1-\lambda_0)U(\delta)}\\ &\overset{(47),(48)}{=} \frac{\lambda_0(1-\lambda_0+\delta\lambda_{\max})^2(Y+O)}{2(1-\lambda_0)(1-\lambda_0-\lambda_{\max})[1+\delta(\lambda_0+\lambda_{\max})]},\end{aligned} \tag{49}$$

$$\begin{aligned}\Psi_n(\delta) &= \frac{\lambda_n(G(\delta)+Y+O)^2}{2(1-\lambda_n)U(\delta)}\\ &\overset{(46),(48)}{=} \frac{\lambda_n(1+\delta\lambda_0-\lambda_{\max})^2(Y+O)}{(1-\lambda_n)(1-\lambda_0-\lambda_{\max})[1+\delta(\lambda_0+\lambda_{\max})]},\end{aligned} \tag{50}$$

对于 $n\in\{1,...,N\}$，以及 $\Phi_n(\delta)=0$ 对所有 $n\geq 2$均成立。

证明。 注意，在一个周期内，$\{\Phi_n\}_{n\in\{1,...,N\}}$ 增加了 $\lambda_0U(\delta)$；如果在某个绿灯时段结束时它严格大于零，那么在前一个周期内它必定减少了 $G(\delta)$。因此，如果 $\{\Phi_n\}_{n\in\{1,...,N\}}$ 在某个绿灯时段结束时严格大于零，那么在前一个绿灯时段结束时它必定减少了

$$\begin{aligned}\Delta_{\Phi_n} &= G(\delta)-\lambda_0U(\delta)\\ &\overset{(46),(48)}{=} \frac{\lambda_0(1+\delta)(Y+O)}{1-\lambda_0-\lambda_{\max}}-\frac{\lambda_0[1+\delta(\lambda_0+\lambda_{\max})](Y+O)}{1-\lambda_0-\lambda_{\max}}\\ &= \lambda_0\delta(Y+O)\\ &> 0.\end{aligned} \tag{51}$$

我们得出结论：对于任意初始状态 $\{\Phi_n\}_{n\in\{1,...,N\}}$，在某个绿灯时段结束时，$\{\Phi_n\}_{n\in\{1,...,N\}}$ 最终必定变为 0，设此时为 t_0。现在注意到，$\{\Phi_n\}_{n\in\{2,...,N\}}$ 只能在绿灯时段内接收交通流（速率至多为单位速率），但也会在绿灯时段内以单位速率疏导交通流，因此从那时起必定始终为零。

Φ_1 从时间 t_0 到时间 t_0+R 以速率 λ_0 增加（δ）$+Y+O$，并从时间 t_0+R（δ）$+Y+O$ 到时间 t_0+R（δ）$+Y+O+\frac{\lambda_0}{1-\lambda_0}$（$R$（$\delta$）$+Y+O$）$=t_0+$以速率~~$1-$~~$\lambda_0$减少

——$\frac{1}{1-\lambda_0}(R(\delta)+Y+O) < t_0+U(\delta)$，使其在该点变为零，并一直保持为零直到 $t_0+U(\delta)$。因此

$$\Phi_1(t_0+t)=\begin{cases}\lambda_0 t, & \text{if } t\in[0,R(\delta)+Y+O],\\ -(1-\lambda_0)t+R(\delta)\\ \quad +Y+O, & \text{if } t\in\left[1,\frac{1}{1-\lambda_0}(R(\delta)+Y+O)\right],\\ 0, & \text{if } t\in\left[\frac{1}{1-\lambda_0}(G(\delta)+Y+O),U(\delta)\right].\end{cases}$$

我们推导得出

$$\begin{aligned}\Phi_1(\delta)&=\lim_{T\to\infty}\frac{1}{T}\int_0^T\Phi_1(t)dt=\frac{1}{U(\delta)}\int_{t=t_0}^{t_0+U(\delta)}\Phi_1(t)dt\\&=\frac{\lambda_0(R(\delta)+Y+O)^2}{2(1-\lambda_0)U(\delta)}\\&\overset{(47)(48)}{=}\frac{\lambda_0(1-\lambda_0+\delta\lambda_{\max})^2(Y+O)}{2(1-\lambda_0)(1-\lambda_0-\lambda_{\max})[1+\delta(\lambda_0+\lambda_{\max})]}.\end{aligned}\tag{52}$$

$$\Phi_1(0)=\frac{\lambda_0(1-\lambda_0)(Y+O)}{2(1-\lambda_0-\lambda_{\max})},\tag{53}$$

当 $\lambda_n=\lambda_1$ 对于 $n\in\{1,...,N\}$ 时，它还有另一种形式

$$\Phi_1(0)=\frac{\lambda_0(1-\lambda_0)(Y+O)}{2(1-\lambda_0-\lambda_1)}.\tag{54}$$

我们还得到，$\{\Psi_n\}_{n\in\{1,...,N\}}$ 在一个周期内增加了 $\lambda_n[U(\delta)-R(\delta)]$，而如果在红灯时段结束时其值严格为正，那么在前一个周期内它必定减少了 $R(\delta)$。

$$\begin{aligned}\Delta_{\Psi_n}&=R(\delta)-\lambda_nU(\delta)\\&\overset{(47)(48)}{=}\frac{\lambda_{\max}(1+\delta)(Y+O)}{1-\lambda_0-\lambda_{\max}}-\frac{\lambda_n[1+\delta(\lambda_0+\lambda_{\max})](Y+O)}{1-\lambda_0-\lambda_{\max}}\\&=\frac{(Y+O)[(1+\delta)\lambda_{\max}-\lambda_n[1+\delta(\lambda_0+\lambda_{\max})]}{1-\lambda_0-\lambda_{\max}}\\&>0.\end{aligned}\tag{55}$$

因此，如果 Ψ_n 在红灯时段结束时严格为正，那么它在前一个红灯时段结束时必定更小。

我们得出结论：对于任意初始状态 Ψ_n，最终都必须在某个红灯时段结束时变为 0，设该时刻为 t_1。然后，Ψ_n 以速率 λ_n 从时间 t_1 增加到时间 $t_1+G(\delta)+Y+O$，再以速率 $1-\lambda_n$ 从时间 $t_1+G(\delta)+Y+O$ 减少到时间 $t_1+G\delta+Y+O+\frac{\lambda_n}{1-\lambda_n}G\delta+Y+O=t_1+()(())\frac{1}{1-\lambda_n}G\delta+Y+O\le t_1+U\delta(())()$，在该时刻变为零，并保持为零，直到 $t_1+U(\delta)$。因此

$$\Psi_n(t_1+t)=\begin{cases}\lambda_n t, & \text{if } t\in[0,G(\delta)+Y+O],\\ -(1-\lambda_n)t+G(\delta)\\ \quad +Y+O, & \text{if } t\in\left[1,\frac{1}{1-\lambda_n}\right](G(\delta)+Y+O),\\ 0, & \text{if } t\in\left[\frac{1}{1-\lambda_n}(G(\delta)+Y+O),U(\delta)\right].\end{cases}\tag{56}$$

我们推知

$$\begin{aligned}\Psi_n(\delta)&=\lim_{T\to\infty}\frac{1}{T}\int_0^T\Psi_n(t)dt=\frac{1}{U(\delta)}\int_{t=t_1}^{t_1+U(\delta)}\Psi_n(t)dt\\&=\frac{\lambda_n(G(\delta)+Y+O)^2}{2(1-\lambda_n)U(\delta)}\\&\overset{(46)(48)}{=}\frac{\lambda_n(1+\delta\lambda_0-\lambda_{\max})^2(Y+O)}{2(1-\lambda_n)(1-\lambda_0-\lambda_{\max})[1+\delta(\lambda_0+\lambda_{\max})]}.\end{aligned}\tag{57}$$

$$\Psi_n(0)=\frac{\lambda_n(1-\lambda_{\max})^2(Y+O)}{2(1-\lambda_n)(1-\lambda_0-\lambda_{\max})},\tag{58}$$

当 $\lambda_n=\lambda_1$ 对于 $n\in\{1,...,N\}$ 时，它还有另一种形式

$$\Psi_n(0)=\frac{\lambda_1(1-\lambda_1)(Y+O)}{2(1-\lambda_0-\lambda_1)}.\tag{59}$$

□

现在，我们利用 IV-D 中的命题，证明在均匀交叉交通流的场景下，“绿波”控制策略的最优性，即对于所有 $n\in\{1,...,N\}$，$\lambda_n\equiv\lambda_1=\lambda_{\max}$。我们以一种情况为例，即 $\lambda_0+2\lambda_1\ge1$、$2\lambda_0+\lambda_1\ge1$，于是得到

$$\Phi_1(\delta)=\frac{\lambda_0(1-\lambda_0+\delta\lambda_1)^2(Y+O)^2}{2(1-\lambda_0)(1-\lambda_0-\lambda_1)^2},$$
$$\Psi_n(\delta)=\frac{\lambda_1(1+\lambda_0\delta-\lambda_1)^2(Y+O)^2}{2(1-\lambda_1)(1-\lambda_0-\lambda_1)^2},$$

其中 $n\in\{1,...,N\}$。此外，$p_1=\frac{\lambda_0}{1-\lambda_0-\lambda_1}$，且 $q_1=1-\lambda_0-\lambda_1$。

(44) 和 (45) 中的下界可化简为

$$\Phi_1\ge\frac{\lambda_0(1-\lambda_1)^2(Y+O)}{2(1-\lambda_0)(1-\lambda_0-\lambda_1)},\tag{60}$$

$$\Psi_n\ge\frac{\lambda_1(1-\lambda_0)^2(Y+O)}{2(1-\lambda_1)(1-\lambda_0-\lambda_1)},\tag{61}$$

除非分别满足 $G_1=\frac{\lambda_0(Y+O)(1-\lambda_0-\lambda_1}{1-\lambda_0-\lambda_1}$ 和 $R_n=\frac{\lambda_1(Y+O)}{1-\lambda_0-\lambda_1}$，否则上述不等式均为严格不等式。

容易验证，当 $\delta\downarrow0$ 时，$\Phi_1(\delta)$ 和 $\Psi_n(\delta)$，$n\in\{1,...,N\}$ 的取值分别趋近于 (60) 和 (61) 中的绝对下界。这表明“绿波”控制策略具有最优性，即不存在其他控制策略能够使任意 Φ_n 或 Ψ_n 取得更低的值（同时保持稳定性）。事实上，在所考虑的场景中，绿波控制策略本质上是唯一的最优策略。

事实上，最优性立即要求 $G_n=G(0)=\frac{\lambda_0(Y+O)}{Y+O}=\frac{(1-\lambda_0-\lambda_1}{()}$、$n\in\{1,...,N\}$，以及 $R_1=R(0)=\frac{\lambda_1}{()}(1-\lambda_0-\lambda_1)$，因为否则式（60）和（61）中的不等式将是严格不等式。此外，出于稳定性的要求，对于所有 $n=2,...,N$，都必须满足 $R_n=R(0)=\frac{\lambda_1(Y+O)}{1-\lambda_0-\lambda_1}$。现在，假设对于某个 $n=2,...,N$，有 $R_n>R(0)$，或者各节点的非绿灯时段并不完全严格同步，并令 W_n 表示在节点 n 开始绿灯时段时，所有节点 $\{1,...,n-1\}$ 的平均已持续同步非绿灯时段。那么，至少存在一个节点 n，使得 $W_n>0$。现在注意到，当所有上游节点 $\{1,...,n-1\}$ 均处于绿灯时段而节点 n 不处于绿灯时段时，Φ_n 必须至少以速率 λ_0 增长。它

随后可以证明 $\Phi_n \geq \frac{\lambda_0 W_n^2}{2\ 1-\lambda_0}$ ()$U_n > 0$，从而破坏最优性。
因此，绿波控制策略的唯一最优性得以证明。