Toll-based Q-learning with non-cooperative agents

Timóteo Fonseca Santos[1], Moisés Gomes de Carvalho[1]

Universidade Federal do Amazonas, Brazil
{tfs, moises}@icomp.ufam.edu.br

Abstract. Congestion is a recurring problem in cities that leads to productivity loss, pollution, and reduced quality of life. Existing traffic congestion resolution techniques are often ineffective or costly. Mathematical analysis and virtual simulation are useful tools to assess the cost-effectiveness of such approaches. Toll-based approaches offer a theoretical foundation for addressing this issue. However, the assumption that all drivers pay tolls may limit real-world efficiency due to non-compliance or economic constraints. This work explores the impacts of different levels of cooperation in toll systems, addressing these challenges. We adapt an existing toll-based approach to handle various scenarios and investigate the feasibility of gradual adoption. Our findings demonstrate that the toll system can be gradually implemented, yielding steady gains and avoiding chaotic behavior, even with non-cooperative agents.

CCS Concepts: • **Computing methodologies** → **Multi-agent reinforcement learning**.

Keywords: machine learning, mct, q-learning, tq-learning, traffic congestion

1. INTRODUÇÃO

Congestionamentos são um problema recorrente em grandes cidades, resultando em perda de produtividade [Somuyiwa et al. 2015], poluição e diminuição na qualidade de vida [Zhong et al. 2017]. As técnicas existentes para resolver congestionamentos no trânsito nem sempre são eficazes ou economicamente viáveis. Por exemplo, a ampliação da capacidade de ruas pode paradoxalmente piorar o fluxo de tráfego [Braess 1968]. No entanto, a implementação de sistemas de pedágio para controlar o fluxo em áreas movimentadas tem demonstrado melhorias observáveis, como evidenciado em Londres [Leape 2006]. Dados os desafios e custos associados à investigação das técnicas de alívio de congestionamento, a análise matemática e a simulação virtual surgem como ferramentas úteis para avaliar o custo-benefício de cada abordagem.

Engarrafamentos podem ser mitigados quando o desempenho ótimo do sistema (SO – *System Optimum*), ou o menor tempo médio de viagem possível (quanto menor esse tempo, melhor o desempenho), diverge do equilíbrio de usuários (UE – *User Equilibrium*), que é quando motoristas não têm como obter maiores ganhos individuais mudando suas estratégias. Nesse caso, há ações que beneficiarão mais o sistema como um todo, mas motoristas que as escolherem sairão no prejuízo. Para convergir o UE ao SO, são necessários incentivos que equilibrem esses possíveis prejuízos. Abordagens com pedágios têm amplo embasamento teórico demonstrando sua eficácia de resolver congestionamentos e são interessantes nesse contexto [Pigou 1920; Hearn and Ramana 1998].

Como veremos nos trabalhos relacionados, uma premissa comum nessas abordagens é que todos os motoristas paguem pedágios. No entanto, a eficiência destas na prática seria limitada pela não-adesão de alguns motoristas, seja por evasão ou limitações econômicas. Além disso, uma vez que a

O presente trabalho foi realizado com apoio da Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Código de Financiamento 001. Este trabalho foi parcialmente financiado pela Fundação de Amparo à Pesquisa do Estado do Amazonas – FAPEAM – por meio do projeto POSGRAD.
A implementação encontra no repositório: `https://github.com/tumut/tql-circumstantial-payment`

instalação de infraestrutura e equipamentos necessários leva tempo, pode ser que a adesão se dê de forma gradual e localizada. E mesmo que as abordagens sejam descentralizadas, alguma centralização pode permanecer necessária para garantir a participação dos motoristas.

Neste trabalho exploramos os impactos que diversos níveis de cooperação podem ter em um sistema de pedágios. Buscamos tratar as seguintes hipóteses: é possível a adaptação de abordagens baseadas em pedágio atuais para lidarem com esses cenários? Sendo possível, os resultados apoiam a adesão gradual aos sistemas de pedágio? Para ambas as questões a resposta é positiva. Criamos uma variação do algoritmo TQ-*learning* [Ramos et al. 2020] de forma a permitir o controle da proporção de motoristas pagantes de pedágios. Com nosso algoritmo realizamos experimentos com múltiplas proporções e apresentamos os resultados. Dessa forma, esperamos expandir o acervo de informações relevantes para tomadas de decisões na resolução prática de engarrafamentos.

A principais contribuições deste trabalho são: 1) uma ferramenta que, diferente das demais encontradas na literatura, permite a investigação dos efeitos de pedágios aplicados apenas parcialmente; 2) uma análise de custo-benefício mais aprofundada para o TQ-*learning*; e 3) evidências de que a aplicação apenas parcial do TQ-*learning* não apresenta grandes riscos para o desempenho do sistema, no geral trazendo melhoras graduais e consistentes.

O restante da obra está organizado da seguinte forma. Na seção 2 analisamos algumas abordagens existentes para lidar com congestionamentos no trânsito, focando especialmente nas técnicas envolvendo pedágios e tarifagem de custo marginal. Na seção 3 apresentamos os conceitos que desempenham um papel fundamental no entendimento dos congestionamentos de tráfego e na avaliação das soluções propostas. Na seção 4 detalhamos o modelo proposto neste trabalho. Na seção 5 mostramos nossa metodologia e apresentamos os resultados obtidos empiricamente. Finalmente, a seção 6 discute os resultados encontrados e conclui o artigo.

2. TRABALHOS RELACIONADOS

Diversas abordagens algorítmicas existem para diminuir o engarrafamento de um sistema, aproximando o UE ao seu SO. O uso de pedágios destaca-se pela sua simplicidade e parcimônia de pressupostos sobre o modelo. É possível calcular *a priori* qual seria o pedágio mínimo ideal de cada rua a fim de obter-se o desempenho ótimo [Hearn and Ramana 1998], mas os cálculos são intensivos por conta do problema ser NP-*hard*. Mesmo com aproximações [Stefanello et al. 2017], ainda requer-se bastante centralização no processo e mudanças no sistema exigem recálculo.

As chamadas recompensas diferenciais (do inglês, *differential rewards*) alteram a recompensa que um agente obtém com base na diferença de desempenho que sua ação causou ao sistema [Wolpert and Tumer 1999]; ou seja, se sua ação piorou o desempenho, a recompensa diminui, e se melhorou, ela aumenta. As recompensas diferenciais têm mais dinamicidade e novamente há aproximações eficazes [Colby et al. 2016], mas seu funcionamento ainda depende de um gerenciamento centralizado e demora-se muitos episódios para obter-se o UE.

O conceito de tarifagem de custo marginal, ou MCT (do inglês, *Marginal-Cost Tolling*) [Pigou 1920], foi uma forma pioneira de resolver o problema de engarrafamentos. Essa abordagem propõe um pedágio cujo custo é proporcional ao quanto o motorista piora o fluxo para os demais usuários. O MCT é o suficiente para garantir a convergência do UE ao SO quando é equivalente à derivada da função de fluxo de uma via [Beckmann et al. 1957]. Dessa forma o MCT permanece relevante até hoje, sendo usado por muitas abordagens. Há trabalhos que exploram o potencial do MCT para o gerenciamento dinâmico dos pedágios, como o Δ-*tolling* [Sharon et al. 2017; Mirzaei et al. 2018], que obtém uma convergência mais rápida do que as recompensas diferenciais. Porém, no geral as abordagens ainda partem do pressuposto de uma agência central com conhecimento global.

O TQ-*learning* apresenta uma abordagem inteiramente descentralizada com convergência melhor

do que nas outras obras [Ramos et al. 2020]. Ele foi expandido como GTQ-*learning* (*Generalized TQ-learning*) para conseguir lidar com preferências heterogêneas de motoristas onde, por exemplo, alguns motoristas não se importam em pagar pedágios maiores para poupar tempo [Ramos et al. 2020]. Porém, ambos ainda pressupõem que todos os motoristas estarão pagando pedágios.

O presente trabalho apresenta uma nova versão do TQ-*learning*, visto que essa técnica é reconhecidamente mais simples e viável para realizar experimentos. Nossa proposta é resultado de uma alteração no algoritmo, permitindo que nem todos os motoristas sejam cooperadores do sistema. Dessa forma, é possível avaliar a tolerância do TQ-*learning* a evasão e adesão parcial.

3. REFERENCIAL TEÓRICO

Aqui apresentamos conceitos essenciais para a compreensão do nosso algoritmo. Primeiro falamos sobre aprendizagem de reforço e Q-*learning*; em seguida, focamos em nosso *baseline*, TQ-*learning*.

3.1 Aprendizagem por reforço e Q-*learning*

Na aprendizagem por reforço um agente aprende por tentativa e erro como se comportar em um dado ambiente [Joshi et al. 1996], não sabendo de antemão quais ações deve realizar mas descobrindo por conta própria quais as que trazem maior recompensa [Sutton and Barto 1998]. Em um ciclo completo, chamado de "episódio", o agente escolhe uma ação e é recompensado por ela. Diversos episódios podem ser realizados em sucessão até o agente chegar num nível de conhecimento satisfatório.

Um algoritmo de aprendizagem por reforço independente de modelos é o Q-*learning* [Watkins 1989]. Ele explora o ambiente a fim de computar uma função $Q : S \times A \to \mathbb{R}$, que retorna a recompensa R estimada por se realizar a ação a no estado s. Ele tem a garantia de convergir a valores ótimos se todos os pares de estado-ação forem experimentados infinitas vezes em um sistema de um único agente [Watkins and Dayan 1992].

Os valores da função são armazenados no que é chamado de tabela-Q e precisam ser atualizados iterativamente. A cada episódio t a nova recompensa encontrada é somada à encontrada anteriormente nas mesmas condições para computar um novo valor na tabela, como na equação 1. Uma taxa de aprendizado $\alpha \in (0, 1]$ controla o quão rapidamente recompensas anteriores são "esquecidas".

$$Q_t(s_t, a_t) \leftarrow (1 - \alpha) \cdot Q_{t-1}(s_{t-1}, a_{t-1}) + \alpha \cdot r_t(s_t, a_t) \tag{1}$$

Diversas heurísticas podem ser usadas no método de escolha do agente. Para impedir que os agentes fiquem presos a um máximo local e assim garantir que todas as ações sejam experimentadas, usa-se a exploração pelo método "ϵ-guloso" (ϵ/epsilon-*greedy*). Nesse heurística um fator de exploração $\epsilon \in [0, 1]$ define a probabilidade do agente escolher uma ação aleatória; caso contrário, ele escolhe a ação com maior recompensa que encontrou até aquele instante [Sutton and Barto 1998, p. 28].

3.2 TQ-*learning*

O TQ-*learning* (do inglês, *toll-based Q-learning*, ou "Q-*learning* baseado em pedágios") [Ramos et al. 2020] combina sistema multiagentes com Q-*learning* e usa MCT para garantir que o UE convirja ao SO. Sendo motoristas os agentes, em um dado sistema de tráfego os motoristas precisam escolher rotas para chegar aos seus destinos, buscando minimizar seus custos de viagem.

O sistema é representado por (G, D, f, τ), sendo $G = (N, L)$ a malha de tráfego com seus nós N e elos L os interligando; D o conjunto de motoristas; f as funções de fluxo; e τ as funções de pedágio. Cada motorista $i \in D$ tem uma origem e um destino, podendo escolher múltiplas rotas para ir de um a outro. Cada elo $l \in L$ tem uma quantidade x_l de motoristas passando por ele e duas funções associadas: uma função de fluxo f_l, que determina o tempo de viagem por ele em função de x_l; e

também uma função de pedágio τ_l, expressa na equação 2 de forma a satisfazer o MCT.

$$\tau_l = x_l \cdot f_l'(x_l) \tag{2}$$

Em cada episódio, os motoristas utilizam a heurística ϵ-*greedy* para escolher rotas como ação. Eles atualizam suas tabelas-Q individuais com base nas recompensas recebidas. As rotas disponíveis para escolha são determinadas a partir dos K caminhos mais curtos na malha de tráfego determinados usando o algoritmo KSP [Yen 1971]. K é um parâmetro.

O fator de exploração ϵ e a taxa de aprendizado α diminuem sistematicamente ao longo do experimento. Essa diminuição gradual torna o comportamento do sistema mais determinístico e menos influenciado por novas descobertas. A diminuição é determinada pelas taxas de decaimento $\lambda, \mu \in]0, 1[$, parâmetros que controlam α e ϵ respectivamente. Em um dado episódio t, $\alpha(t) = \lambda^t$ e $\epsilon(t) = \mu^t$.

O custo de se atravessar um elo é o tempo de viagem somado ao pedágio (equação 3), e o custo da rota $a_{i,t}$ escolhida pelo motorista i no episódio t é o somatório dos custos dos seus elos (equação 4). A recompensa r pela rota escolhida é inversamente equivalente ao custo (equação 5). Assim, ao maximizar a recompensa, os agentes estarão também minimizando o custo.

$$c_l(x_l) = f_l(x_l) + \tau_l(x_l) \tag{3}$$

$$C_{a_{i,t}} = \sum_{l \in a_{i,t}} c_l \tag{4}$$

$$r(a_{i,t}) = -C_{a_{i,t}} \tag{5}$$

4. ALGORITMO

Propomos um TQ-*learning* com agentes não-cooperativos, onde motoristas (os agentes) podem ou não serem usuários do sistema de pedágios e apenas pagam caso sejam usuários. Introduzimos um parâmetro $\upsilon \in [0, 1]$ que controla a proporção de motoristas que são usuários, sendo 0 o caso em que ninguém paga pedágio (o que tende ao UE) e 1 o caso em que todos pagam (tendendo ao SO). Valores intermediários de υ possibilitam que usuários e não-usuários interajam no mesmo sistema de tráfego.

A cada motorista $i \in D$ foi associado um valor booleano $s_i \in \mathbb{B}$ que determina se o motorista é usuário (caso seja verdadeiro) ou não (caso $\neg s_i$). Esse valor é definido para cada motorista no começo da execução com probabilidade υ de ser verdadeiro, e permanece fixo até o fim da execução.

As funções de custo do TQ-*learning* foram alteradas nas equações 6 e 7 a fim de levar em consideração o s_i. A notação $[s_i]$, chamada de colchetes de Iverson [Knuth 1992], converte s_i em 1 caso seja verdadeiro e em 0 caso seja falso. Dessa forma, o pedágio é zerado caso o motorista não seja usuário.

$$c_{l,i} = f_l(x_l) + [s_i] \cdot \tau_l(x_l) \tag{6}$$

$$C_{i,a_{i,t}} = \sum_{l \in a_{i,t}} c_{l,i} \tag{7}$$

O algoritmo 1 formaliza o funcionamento do nosso TQ-*learning* com agentes não-cooperativos. A cada episódio t, cada motorista i escolhe uma rota $a_{i,t}$. Se ele for usuário, ele pagará pedágios como parte do seu custo de viagem. Ele é recompensado de acordo com o quanto minimizou seus custos; e assim sucessivamente. A média de tempos de viagens de motoristas no último episódio da simulação, denotada por ν, servirá como a medida de desempenho do sistema. Quanto menor o ν, melhor.

5. EXPERIMENTOS

Apresentamos aqui os resultados empíricos de experimentos com diferentes níveis de cooperação entre os motoristas, mostrando como o fator υ influencia o ν das simulações e de que maneira isso ocorre.

Algoritmo 1 TQ-*learning* com agentes não-cooperativos

function EXPERIMENTO($P; T; K; \lambda; \mu; \upsilon$)
  $(G, D, f, \tau) := P; (N, L) := G;$
  $A \leftarrow$ inicialização de opções de rotas em função de K;
  $s_i \leftarrow \mathcal{U}[0, 1) < \upsilon, \forall i \in D;$ ▷ determina quem será ou não usuário
  $Q_{i,t}(a) \leftarrow 0, \forall i \in D, \forall a \in A_i, \forall t \in [0..T];$ ▷ inicialização da tabela-Q
  for $t \in [1..T]$ **do**
    $\alpha \leftarrow \lambda^t; \epsilon \leftarrow \mu^t;$ ▷ atualiza-se a taxa de aprendizado e o fator de exploração
    $x_l \leftarrow 0, \forall l \in L;$ ▷ fluxos são reinicializados
    for $i \in D$ **do**
      $a_{i,t} \leftarrow$ escolha de rota com ϵ-*greedy*;
      $x_l \leftarrow x_l + 1, \forall l \in a_{i,t};$ ▷ incrementa-se os fluxos de cada elo l na rota $a_{i,t}$
    $f \leftarrow$ computa o tempo de viagem para todos os elos e rotas;
    for $i \in D$ **do**
      $f_{a_{i,t}} \leftarrow$ observa o tempo de viagem na rota $a_{i,t}$;
      $\tau_{a_{i,t}} \leftarrow$ calcula o pedágio pra rota;
      $r(a_{i,t}) \leftarrow -(f_{a_{i,t}} + [s_i] \cdot \tau_{a_{i,t}});$ ▷ recompensa de i; pedágio condicionado por s_i
      $Q_{i,t}(a_{i,t}) \leftarrow (1 - \alpha) \cdot Q_{i,t-1}(a_{i,t-1}) + \alpha \cdot r(a_{i,t});$ ▷ atualiza tabela-Q de i

5.1 Metodologia

Os experimentos foram realizados usando as mesmas malhas de tráfego sintéticas usados no trabalho original do TQ-*learning* [Ramos et al. 2020, p. 15][1], onde são descritas. Cada malha usa um valor K específico. As malhas (com seus respectivos valores K) são: B^1 (3), B^2 (5), B^3 (7), B^4 (9), B^5 (11), B^6 (13), B^7 (15), BB^1 (3), BB^3 (8), BB^5 (4), BB^7 (4) e OW (8).

Em todos os casos são usados os valores $\lambda = \mu = 0.99$ para as taxas de decaimento e $T = 1000$ para a quantidade de episódios, parâmetros que segundo [Ramos 2018, p. 114] produzem os melhores resultados para obtenção do SO nas malhas sintéticas. Variamos cada malha ao longo de $\upsilon \in \{0, 0.25, 0.5, 0.75, 1\}$. Para cada permutação de parâmetros os experimentos são repetidos 30 vezes, daí agrupamos os valores v e tiramos sua média e desvio padrão.

É esperado que os casos $\upsilon = 0$ e $\upsilon = 1$ sejam próximos do UE e do SO do sistema, respectivamente. Comparamos os resultados com valores de UE e SO obtidos na literatura [Ramos 2018, p. 113], usando o cálculo de proximidade na equação 8 (valores de referência sendo representados como v^*).

$$\phi(v; v^*) = 1 - \frac{|v - v^*|}{v^*} \tag{8}$$

5.2 Resultados

A Tabela I apresenta as médias dos valores de v obtidos para os experimentos ao logo das permutações de υ, para cada rede simulada. O desvio padrão pode ser observado entre parênteses.

O caso $\upsilon = 0$, em que nenhum motorista paga pedágio, tende ao UE. O caso $\upsilon = 1$, em que todos pagam, ao SO. No geral, trouxeram o pior e o melhor desempenho de cada rede, respectivamente. Isso significa que os casos intermediários ($0 < \upsilon < 1$) produziram, em quase todos os casos, resultados entre os v dos casos extremos ($\upsilon \in \{0, 1\}$). A única exceção a esse padrão foi a rede OW, cujos casos intermediários apresentaram resultados melhores do que SO. De todo modo, em nenhum momento um $\upsilon > 0$ resultou em desempenho pior do que o de $\upsilon = 0$ na mesma rede.

[1] As malhas de tráfego foram obtidas do repositório público disponível em `https://github.com/goramos/transportation_networks`

	υ				
Rede	0 (∼UE)	0.25	0.5	0.75	1 (∼SO)
B^1	18.470 (0.606)	17.805 (0.059)	16.257 (0.032)	15.309 (0.014)	15.000 (5.815×10^{-5})
B^2	27.652 (1.156)	27.472 (0.372)	25.828 (0.052)	24.381 (0.030)	23.334 (0.004)
B^3	37.397 (1.138)	36.938 (0.761)	35.600 (0.124)	33.907 (0.038)	32.500 (3.519×10^{-4})
B^4	47.053 (1.158)	46.334 (0.867)	45.156 (0.345)	43.588 (0.053)	42.001 (5.387×10^{-4})
B^5	56.322 (0.889)	55.605 (0.819)	54.714 (0.335)	53.288 (0.090)	51.668 (9.758×10^{-4})
B^6	66.231 (0.682)	65.406 (0.466)	64.436 (0.257)	63.014 (0.097)	61.430 (0.001)
B^7	75.961 (0.343)	75.130 (0.297)	74.070 (0.146)	72.748 (0.091)	71.255 (0.006)
BB^1	10.000 (0.000)	8.904 (0.040)	8.126 (0.026)	7.655 (0.012)	7.500 (0.000)
BB^3	22.009 (0.213)	21.402 (0.104)	20.453 (0.035)	19.602 (0.025)	19.000 (7.824×10^{-4})
BB^5	50.561 (0.039)	49.390 (0.043)	48.415 (0.035)	47.601 (0.027)	47.003 (0.002)
BB^7	124.157 (0.140)	122.965 (0.071)	121.954 (0.056)	121.170 (0.068)	120.541 (0.049)
OW	67.199 (0.010)	66.971 (0.003)	66.969 (0.003)	66.973 (0.005)	66.987 (0.006)

Table I. Tempos médios de viagem v observados (com o desvio padrão) para cada valor de υ.

Rede	Ref. UE	Prox. UE	Ref. SO	Prox. SO
B^1	20	0.92348 (0.03)	15	1.00000 (3.88×10^{-6})
B^2	30	0.92172 (0.04)	23.3333	0.99995 (1.78×10^{-4})
B^3	40	0.93491 (0.03)	32.5	0.99999 (1.08×10^{-5})
B^4	50	0.94106 (0.02)	42	0.99998 (1.28×10^{-5})
B^5	60	0.93869 (0.01)	51.6667	0.99998 (1.89×10^{-5})
B^6	70	0.94616 (9.74×10^{-3})	61.43	0.99999 (1.38×10^{-5})
B^7	80	0.94952 (4.29×10^{-3})	71.25	0.99993 (8.81×10^{-5})
BB^1	10	1.00000 (0.00)	7.5	1.00000 (0.00)
BB^3	22	0.99231 (5.69×10^{-3})	19	0.99999 (4.12×10^{-5})
BB^5	50.3	0.99481 (7.79×10^{-4})	47	0.99993 (4.30×10^{-5})
BB^7	123.84	0.99744 (1.13×10^{-3})	120.5	0.99966 (4.06×10^{-4})
OW	67.16	0.99942 (1.50×10^{-4})	66.92	0.99900 (9.58×10^{-5})
Média		0.96163 (0.04)		0.99987 (3.07×10^{-4})

Table II. Proximidades (com desvio padrão) dos resultados obtidos para os valores de UE e SO de referência.

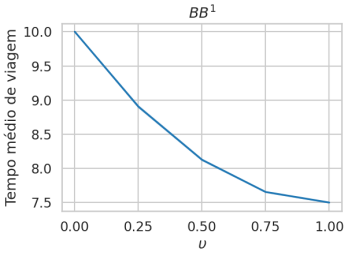

Fig. 1. BB^1, exemplo de curva "côncava".

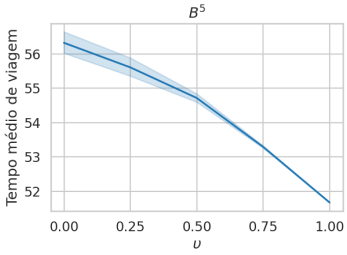

Fig. 2. B^5, exemplo de curva "convexa";

A Tabela II mostra as proximidades entre os valores de UE e SO e seus respectivos valores de referência. No final encontra-se as médias de proximidade considerando todas as redes em conjunto.

Houve correspondência praticamente exata com as proximidades encontradas para os valores de SO em [Ramos 2018, p. 115]. No entanto, os valores de UE não tiveram uma proximidade tão boa devido ao uso de parâmetros otimizados para o cálculo do SO, resultando em desempenhos ligeiramente melhores. Vale ressaltar que a rede BB^1 apresentou cálculos exatos para seus UE e SO, inclusive obtendo variância 0.

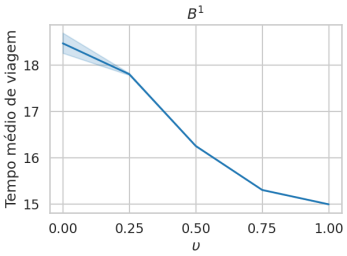

Fig. 3. B^1, exemplo de curva sinuosa.

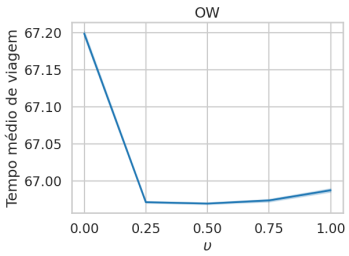

Fig. 4. OW, o *outlier*.

Ao formatar os resultados em gráficos lineares quatro grupos emergiram com base no formato da linha, exemplificados por suas redes mais flagrantes nas Figuras 1, 2, 3 e 4. As áreas sombreadas em torno das linhas representam o intervalo de confiança de 95%.

A Figura 1 representa as redes com melhorias "côncavas" (constituídas por BB^1, BB^5 e BB^7), porque o Δv diminui à medida que υ se aproxima de 1, sendo maior no começo do que no final. Isso significa dizer que, no geral, a diferença entre valores de v é maior em intervalos de υ próximos de 0 do que de 1. São, portanto, redes mais vulneráveis ao efeito de rendimentos decrescentes. A Figura 2 representa as redes com melhorias "convexas" (B^4, B^5, B^6 e B^7), cujo Δv aumenta à medida que υ se aproxima de 1. A Figura 3 representa as redes cujas melhorias realizam uma curva sinuosa (B^1, B^2, B^3 e BB^3), tendo valores de Δv que começam pequenos, aumentam em torno de $\upsilon = 0.5$, e diminuem de novo logo em seguida. As redes encontram-se distribuídas ao longo desses três grupos de maneira aproximadamente equilibrada. A rede OW, na Figura 4, é um *outlier*, com o melhor valor v ocorrendo em $\upsilon = 0.5$ e regredindo a partir daí. Em nenhuma outra rede houve regressão do v em si.

6. CONCLUSÃO E TRABALHOS FUTUROS

Neste artigo, apresentamos os resultados de experimentos com uma nova versão do algoritmo TQ-*learning*, cujo diferencial é permitir controlar a proporção de motoristas isentos de pedágio. Nosso objetivo foi avaliar o desempenho do algoritmo em diferentes níveis de adesão ao sistema de pedágios e suas implicações para implementação prática, proporcionando uma análise mais aprofundada do custo-benefício do TQ-*learning*. Investigamos duas principais questões: 1) os riscos da adesão parcial trazer mais prejuízos do que nenhuma adesão e 2) a necessidade de adesão total aos pedágios para obter resultados satisfatórios.

A ausência de valores v piores do que o UE é evidência de que a aplicação parcial do sistema de pedágios não prejudicaria o fluxo de trânsito, contradizendo a ideia de que seria pior aplicar pedágios de forma parcial do que não aplicá-los. Além disso, observa-se melhorias de desempenho mesmo em proporções reduzidas, não sendo necessário atingir uma "massa crítica" de usuários para obter resultados. As redes que apresentaram curvas convexas e sinuosas (ver Figuras 2 e 3) foram as que mais chegaram perto de mostrar alguma estagnação inicial, porém sem grande significância.

A rede OW não convergiu tão bem aos seus valores de UE e SO de referência quanto as demais redes. Isso indica que ela talvez precise de mais episódios para convergência. Observamos que em todos os casos intermediários seu desempenho foi melhor do que no $\upsilon = 1$. É improvável que isso seja porque de fato a proporção parcial traz melhor desempenho do que a adesão total, então pode ser um indicador de que a simulação converge mais rápido nos casos de proporção parcial.

Tudo indica que a adesão ao sistema de pedágios pode ser feita de maneira gradual. Não é necessário que ele seja implementado de uma só vez para obtermos resultados. Ele aparenta ser relativamente robusto à não-cooperação de pelo menos alguns agentes, e em nenhum caso os efeitos da não-cooperação introduzem um comportamento caótico ao sistema. As melhorias são graduais e, no geral, consistentes.

Para trabalhos futuros, vale a pena investigar possíveis fórmulas descrevendo a relação de ν com υ de maneira geral. Experimentos com uma maior resolução de valores pra υ seriam úteis nesse caso e permitiriam (por exemplo) a análise por regressão linear. Além disso, mais episódios e alterações nos parâmetros poderiam ser experimentados com a rede OW para avaliar a hipótese de que o sistema precisa de mais tempo para convergir. Por último, as redes utilizadas são todas sintéticas, e temos a intenção de realizar experimentos com redes de ruas obtidas de dados reais.

REFERENCES

Beckmann, M., McGuire, C. B., Winsten, C. B., and Koopmans, T. C. Studies in the economics of transportation. *The Economic Journal* 67 (265): 116–118, 1957.

Braess, D. Über ein paradoxon aus der verkehrsplanung. *Unternehmensforschung Operations Research - Recherche Opérationnelle* vol. 12, pp. 258–268, 12, 1968.

Colby, M., Duchow-Pressley, T., Chung, J. J., and Tumer, K. Local approximation of difference evaluation functions. In *Proceedings of the 2016 International Conference on Autonomous Agents & Multiagent Systems*. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 521–529, 2016.

Hearn, D. W. and Ramana, M. V. pp. 109–124. In P. Marcotte and S. Nguyen (Eds.), *Solving Congestion Toll Pricing Models*. Springer US, Boston, MA, pp. 109–124, 1998.

Joshi, D. J., Kale, I., Gandewar, S., Korate, O., Patwari, D., and Patil, S. Reinforcement learning: A survey. *Journal of Artificial Intelligence Research* vol. 1311 AISC, pp. 297–308, 1996.

Knuth, D. E. Two notes on notation, 1992.

Leape, J. The london congestion charge. *Journal of Economic Perspectives* vol. 20, pp. 157–176, 9, 2006.

Mirzaei, H., Sharon, G., Boyles, S., Givargis, T., and Stone, P. Enhanced delta-tolling: Traffic optimization via policy gradient reinforcement learning. In *2018 21st International Conference on Intelligent Transportation Systems (ITSC)*. pp. 47–52, 2018.

Pigou, A. C. *The Economics of Welfare*. Routledge, 1920.

Ramos, G. de. O. *Regret Minimisation and System-Efficiency in Route Choice*. Ph.D. thesis, Universidade Federal do Rio Grande do Sul, Brazil, 2018.

Ramos, G. de. O., da Silva, B. C., Rădulescu, R., Bazzan, A. L. C., and Nowé, A. Toll-based reinforcement learning for efficient equilibria in route choice. *Knowledge Engineering Review*, 2020.

Ramos, G. de. O., Rădulescu, R., Nowé, A., and Tavares, A. R. Toll-based learning for minimising congestion under heterogeneous preferences. In *Proc. of the 19th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2020)*, B. An, N. Yorke-Smith, A. El Fallah Seghrouchni, and G. Sukthankar (Eds.). IFAAMAS, Auckland, New Zealand, pp. 1098–1106, 2020.

Sharon, G., Hanna, J. P., Rambha, T., Levin, M. W., Albert, M., Boyles, S. D., and Stone, P. Real-time adaptive tolling scheme for optimized social welfare in traffic networks. AAMAS '17. International Foundation for Autonomous Agents and Multiagent Systems, Richland, SC, pp. 828–836, 2017.

Somuyiwa, A. O., Fadare, S. O., and Ayantoyinbo, B. B. Analysis of the cost of traffic congestion on worker's productivity in a mega city of a developing economy. *International Review of Management and Business Research* 4 (3): 644, 2015.

Stefanello, F., Buriol, L. S., Hirsch, M. J., Pardalos, P. M., Querido, T., Resende, M. G. C., and Ritt, M. On the minimization of traffic congestion in road networks with tolls. *Annals of Operations Research* 249 (1): 119–139, Feb, 2017.

Sutton, R. S. and Barto, A. G. Reinforcement learning: An introduction. *IEEE Transactions on Neural Networks* 9 (5): 1054–1054, 1998.

Watkins, C. J. C. H. *Learning from Delayed Rewards*. Cambridge University, 1989.

Watkins, C. J. C. H. and Dayan, P. Q-learning. *Machine Learning 1992 8:3* vol. 8, pp. 279–292, 5, 1992.

Wolpert, D. H. and Tumer, K. An introduction to collective intelligence. *CoRR* vol. cs.LG/9908014, 1999.

Yen, J. Y. Finding the k shortest loopless paths in a network. *Management Science* 17 (11): 712–716, 1971.

Zhong, N., Cao, J., and Wang, Y. Traffic congestion, ambient air pollution, and health: Evidence from driving restrictions in beijing. *Journal of the Association of Environmental and Resource Economists* 4 (3): 821–856, 2017.