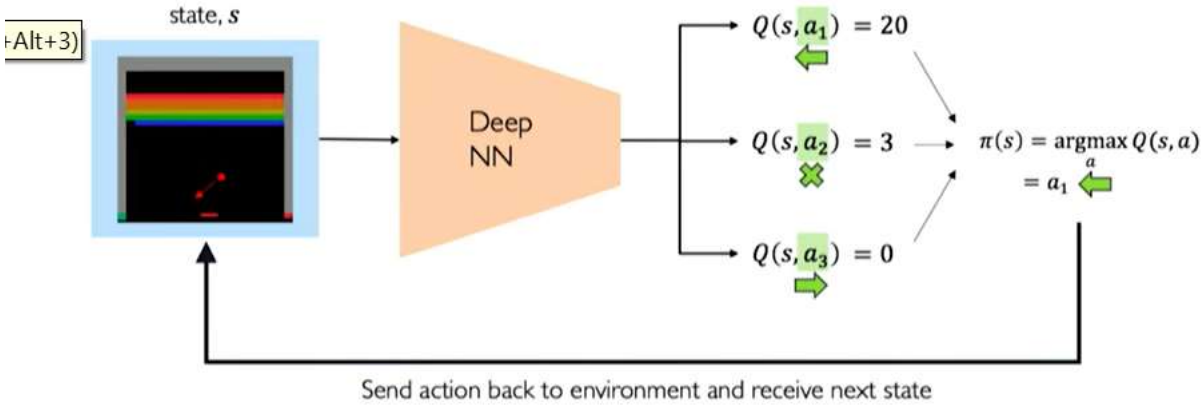
Uma palavra sobre Reinforcement Learning (RL)

De uma forma geral o RL busca tomar a melhor ação a ser tomada diante de um Estado apresentado pelo sistema. Para isso ele precisa de uma função Q (Q-Function), que mapeia as duplas (ações estado) no total de recompensas até o final do funcionamento do sistema.

a) **Se a Q-Function é conhecida**

Uma das maneiras de se obter Q é por DeepLearning, em que a rede recebe como entrada o estado (s_t) . Nesse estado, a rede deve aprender a tomar a_{it} ação que fornece o maior valor Q, ou seja, a_{it} = a_i tal que Q(st, ai) é máximo, para 1<i<n, em que n é o nímero de ações possíveis. No processo de treinamento a rede é retroalimentada com a ação e o próximo estado, conforme ilustra a figura a seguir:



No caso da figura acima, o estado é definido pelos *pixels* da teça de um jogo, por exemplo, e as ações ão aquelas que um jogador pode tomar

Cabe ressaltar algumas ressalvas , expressas em 32.25 do vídeo (https://www.youtube.com/watch?v=-WbN61qtTGQ):

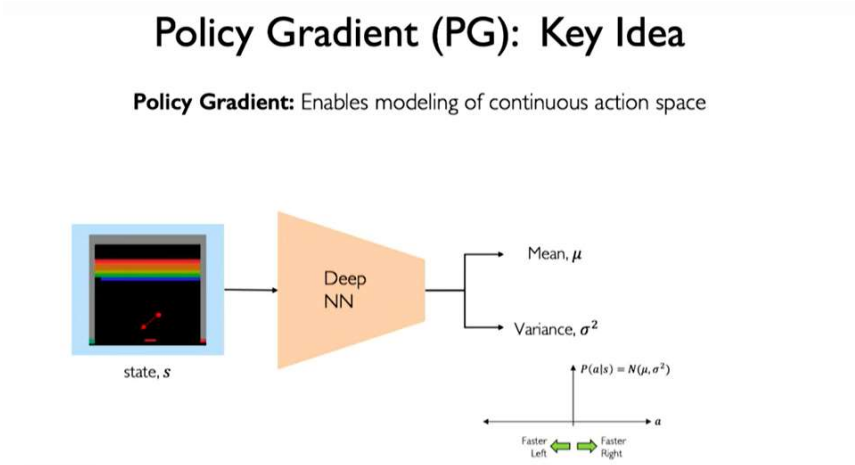
- Essa técnica é para conjunto de cenários discretos

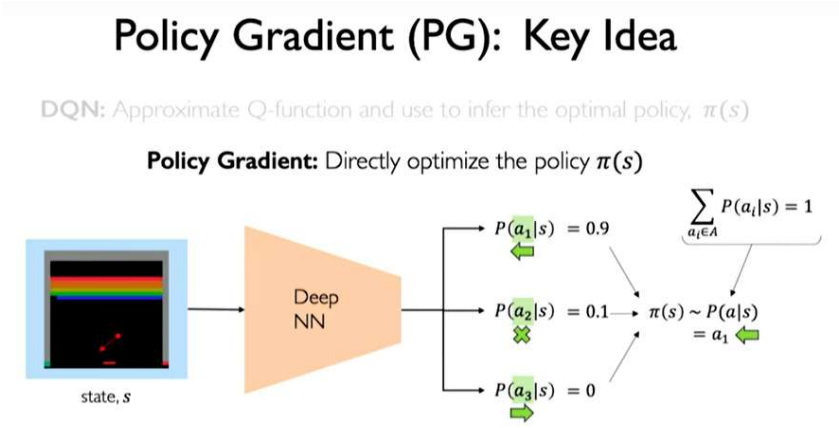
- Também é determinística um Estado → uma ação.

Para contornar esses problemas, se utiliza o Policy Gradient visto a seguir:

b) Policy Gradient

Agora, em vez de se considerar a Q, considera-se a P(a|s), ou seja, qual é a probabilidade de se obter a melhor recompensa, tomando a ação a no estado s. Sendo assim a rede será treinada para obter uma média e uma variância, que melhor expresse os valores de Q e assim obtenha as probabilidades. As figuras abaixam ilustam o processo





Utilização de LSTM no contexto do RL, ou seja para mapear estados em ações que deram os melhores resultados, o que é diferente da nossa abordagem!

#	Ref	Artigo	Pontos Principais	Diferencial
1	**	Eagle: Refining Congestion Control by Learning from the Experts	-Utiliza LSTM, mas no contexto de um DRL, tradicional. -Se vale do BBR para “momentos críticos” -Procura ajustar efetivamente a janela de congestionamento -Divisão do modelo de treinamento em (startup, queue draining, and bandwidth probing) -Protocolo Reativo	-Utilização da LSTM Pura -Sem utilização de um esquema pré-existentes **-Previsão da situação da rede passos a frente (proativo)** **-A questão de se considerar condições prévias da rede é recorrente.** **-Nossa vantagem também seria a possibilidade de se dispensar a tabela de mapeamento. Em contra partida com um histórico de 10 [23] chegou a bons resultados.**
2	01	Classic Meets Modern: a Pragmatic Learning-Based Congestion Control for the Internet	-Divide o controle em alto-nível, treinado por DRL (Deep Reinforcement learning), que interfere na cwnd do baixo nível -Apresentou bons resultados Olhar para os CC por aprendizado puro (Aurora, Indigo, Remy e Vivace) -Aurora também é por RL	
3	06	Learning-TCP: A stochastic approach for efficient update in TCP congestion window in ad hoc wireless networks	-Baseia-se no intervalo entre as chegadas de ACK -Redução pró-ativa -5 ações levam a baixo refinamento, mais um problema em RL. -Então é introduzido um aprendizado que atualiza a distribuição de probabilidade de se tomar uma determinada ação, que	Mostra que estamos no caminho certo. Mais um problema e RL: O espaço entre as ações. O módulo à parte pode ser interessante para nós

			consiste em um incremento/decremento na cwnd. -implementa o módulo de aprendizagem à parte Another problem is that the meann sharply rises to a high value and remains over there for a long-lasting congestion in the network. However, in order to learn the network conditions better, the proposed learning mechanism requires meann when there is no congestion in the network. That is, meank rises (or falls) when there is a sudden increase (or decrease) in the congestion in the network. When we take the ratio of these two means (i.e., meank meann), the ratio will be much above to the value of 1 when the congestion is rising. So, when this ratio is much below to the value of 1, we can infer minimal or no congestion in the network. Therefore, we update safeMean with meann when this ratio is less than a threshold 0.75; further we take the values for k and n as 20 and 100, respectively, which are derived via empirical studies. We validate the updating of safeMean with	Procurar fazer o mesmo em nosso modelo. Seria uma excelentee ideia para se evitar a necessidade do valor do preenchimento do buffer do roteador….

			meann in Section 5.	
4	23	A Deep Reinforcement Learning Perspective on Internet Congestion Control (Aurora)	-Também baseado em Reinforcement Learning -	
5	24	A congestion control method of SDN data center based on reinforcement learning	Foco em data centers network (DCN), que tiveram suas demandas de dados ampliadas principalmente por causa do big data. SDN – Separação do plano de controle do plano de dados	Sem SDN.
6	27	A Deep Reinforcement Learning based Congestion Control Mechanism for NDN	-Trabalha com NDN, um paradigma completamente das redes TCP/IP Traz a possibilidade de crashar e recorre ao tradicional	TCP-IP
7	28	QTCP: Adaptive Congestion Control with Reinforcement Learning	-Reforça a ideia de adaptação da camada de transporte -Traz a ideia de um aprendizado online -Artigo extremamente alinhado com a ideia geral proposta há muito tempo, nos primórdios do MTO! Se utiliza das mesmas features (avg ack, avg send, avg rtt) para determinar o estado da rede Muito semelhante a nossa abordagem, mas com RL, sem falar em previsão de filas, foca no ajuste do cwnd, a partir das	our learning process can be interrupted in certain situations that are handled by standard TCP mechanism such as slow start and fast recovery A princípio não haverá no nosso!

			mesmas variáves. Nos experimentos usa dumbel com 120 ms de RTT e um CC, o NewReno (bem mais simples do que a nossa).	
8	29	Learning-based and Data-driven TCP Design for Memory-constrained IoT	Mais uma vez RL, com Q-learning Both Remy and PCC regard the network as a black box and focus on looking for the change in the sending rate that can lead to the best performance, without directly interpreting the environment or making use of previous experience. Trecho identico no artigo anterior!!! Observar a questão do PCC -Mais uma vez, discretização do espaço estado ação.	Noso seria ML, com foco nos roteadores, com a consolidação de um modelo. Há necessidade de se avaliar se haverá adaptação o se o modelo será carregado.
9	32	Self-learning Congestion Control of MPTCP in Satellites Communications	-Foco no MPTCP -Visando estruturas de satélites low earth orbit(LEO) 500 a 1500 km -O desafio é como fazer o controle de congestionamento de múltiplos MPTCP. -Os trabalhos propostos até agora, sofrem dos mesmos problemas de muitos do singlepath -Mais uma vez reinforcement learning -Nada sobre datamining	-Nada sobre fila nos roteadores -Nada sobre previsão passos à frente.

			-Pouquíssimos gráficos	
10	42	Dynamic TCP Initial Windows and Congestion Control Schemes Through Reinforcement Learning	-Agora o foco é short-flows, para os quais não há tempo de convergir -Inicial window = 10, exatamente conforme ns3 -TCP RL – Tenta otimizar para fluxos longos e curtos -Se baseia nas observações de um webservice no servidor	-De posse de um modelo treinado, poderíamos já levar a janela para um ponto ótimo, desde o início da transmissão.
11	Extra0001	RNN-based Approach to TCP throughput prediction	-Presença de um módulo coletor de métricas que alimenta uma RNN a **cada perda** **-Não foi identificado como foi realizado o datamining** **Não foi identifcada a maneira como foi treinada a RNN**	Nosso diferencial seria ser pró-ativo.
12	Extra004	Forecasting TCP’s Rate to Speed up Slow Start	-Calcado nas medições do BBR de banda do BBR para teinar a rede neural -Focado no slow-Start, que é onde se estima a banda…. Depois de treinar offline, o modelo é incorporado à Pilha TCP - A ideia é fornecer antecipadamente a estimativa de banda do BBR no solwstart, o que pode minimizar o probe;	-Nossa proposta é diferente pois pocura focar na predição como um todo. É interessante obsservar a quantidade de épocas (300) e o quanto é calcado no bbr, utilizando sua estimativa, -O foco naa fila do roteador dá uma visão mais ampla da rede como um todoi
13		Learning-TCP: A Novel Learning Automata Based Reliable Transport Protocol for Ad hoc Wireless Networks	Foco em redes wireless Se adaptar às condições da rede por meio de ACK e DUPACK	

14		Experience-Driven Congestion Control: When Multi-Path TCP Meets Deep Reinforcement Learning	Mais uma vez DRL MPTCP O agente vai controlar todos os fluxos de um MCTCP A ação deriva da saída da LSTM, que aprende os padões da REDE epoch=50.000 e learning Rate = 0.0001 and 0.001 Due to the light weight of our design, there is no need for any special device (such as GPU) for training	
15		ICTCP: Incast Congestion Control for TCP in Data Center Networks	Para sanar o TCP-Incast, que é um congestionamento que ocorre quando vários servidores fazem requisição em uma rede de alta velocidade e baixa latência, tal como em Data Centers Ajuste da rwnd proatiamente, antes da perda Só para redes DataCenters like Quando o a taxa de recepção é muito baixa, diminui a rwnd dos fluxos dento do DataCnter	
16		Reinforcement Learning-Based Neural Network Congestion Controller for ATM Networks	Em geral, mesmos princípios do RL	