Let's play congestion control*

Extended Abstract†

Maximilian Bachl
Institute of Telecommunications
Gußhausstraße 25
1040, Vienna, Austria
maximilian.bachl@tuwien.ac.at

Tanja Zseby
Institute of Telecommunications
Gußhausstraße 25
1040, Vienna, Austria
tanja.zseby@tuwien.ac.at

Joachim Fabini
Institute of Telecommunications
Gußhausstraße 25
1040, Vienna, Austria
joachim.fabini@tuwien.ac.at

ABSTRACT

This paper proposes a novel method of Congestion Control that uses on-line reinforcement learning (RL) to maintain the optimum congestion window with respect to a given utility function.

To this end we use a neural network based approach that can be initialized with a previously trained neural network or with random weights (without any prior knowledge). As in congestion control the rewards used in the learning process naturally come delayed and bit by bit, we propose a novel formulation of reinforcement learning that allows us to deal with delayed and partial rewards.

We show that our method converges to a stable, close-to-optimum solution within the order of minutes and then commonly outperforms existing congestion control algorithms in typical networks. Thus, for the first time, we demonstrate that ANN based reinforcement learning without any preknowledge can feasibly be done on-line and can compete with hand-crafted solutions given long enough episodes.

CCS CONCEPTS

•**Networks** →**Transport protocols;** •**Theory of computation** →**Multi-agent reinforcement learning;** •**Computer systems organization** →*Neural networks;*

KEYWORDS

congestion control, machine learning, reinforcement learning, artificial neural networks

1 INTRODUCTION

Recent advances in neural network based reinforcement learning (RL) for the purpose of playing video games raise the question if it is also possible to formulate sending Internet data in a fashion similar to a video game. The rules of the "Game of Congestion Control" can be summarized approximately as follows: Each player wants to send (1) as many packets as possible (2) with as little delay as possible (3) while losing as few packets as possible. As, for instance, for some applications overall throughput is more important than little packet loss, each application can define its preference using a custom utility function. Clearly, it is also desirable from an overall point of view that all participating players play the game in a way that does not give them an unfair advantage over others. When several players (from now on called senders) share an Internet link under the same conditions, they should receive an equal share of the reward.

When performing Congestion Control (CG) TODO: What should be capitalized and what shouldn't? Capitalize RL, CG, ANN?, each sender uses a set of observed environment conditions to determine what action he should take to maximize his reward in the future. The environment conditions can be any metrics that the sender can obtain from the network, for example the mean round-trip time of the last received packets, the packet loss rate etc. An action is a change to the sending rate: For example, if a sender perceives an increase in the loss rate it might be advisable to lower the sending rate.

From this formulation of Congestion Control, one can observe a couple of unique features of the problem that seem to prohibit the use of RL:

- When playing a video game, if one pushes a button, one can see the consequences of an action immediately. However, in Congestion Control, the rewards are always **delayed** by one round-trip time.
- Due to the delay, by the time an action receives its reward, probably other actions have already been carried out in the mean time. Thus, contrary to classical RL, **action and rewards are out of sync**.
- **Actions and rewards are not atomic**: If an action causes three packets to be sent, we consider these three packets being sent three *partial actions*. In case these packets are transmitted correctly, the receiver will send three acknowledgements. Each of these acknowledgements is a *partial reward* that already unveils new information on the current state of the network and enables the sender to take an action based on this new information. However, only if all partial rewards have been received, the sender can assemble the overall reward of the corresponding action and can perform an update of the underlying RL logic.

To address these problems, we develop a new formulation of RL called *Partial Action Learning* (PAL). PAL is a superset of reinforcement learning: If one uses PAL for a learning problem without delay, asynchronicity and partial actions/partial rewards, one gets classical RL.

While using PAL to train an optimum congestion control for a specific range of network scenarios in an off-line fashion is possible, it is something that has already been done previously in an approach called *Remy* [11]. The only seizable advantage that PAL could provide here is increased training speed. Thus we want to show that it is not only possible to learn congestion control by using machine learning but that it is even possible to do so on-line and without any preknowledge about the network environment.

Using a utility function that encourages high throughput and discourages packet loss (adapted from [3]) we see that our machine learning based approach can learn congestion control that maximizes this objective given a couple of minutes of time (without any pre-training) 1, which – to our knowledge – is the first demonstration that neural network based on-line reinforcement learning is possible.

After spending up to a few minutes with suboptimal performance during initial exploration (if no online pre-training was used), our approach can compete with established algorithms and achieve superior throughput with packet loss at an average of less than 5%, while still gradually improving and adapting its way of performing congestion control.

2 RELATED WORK

2.1 Congestion Control

Congestion control has been implemented in TCP since the 1980s as a series of congestion collapses dramatically decreased applications' throughput in the Internet, which lead to the development of the Tahoe and Reno algorithms [7]. These approaches as well as most others maintain a congestion window which stands for the maximum amount of data that are allowed to be traveling in the network without having been acknowledged by the receiver. In general, receiving acknowledgements of data that were successfully transmitted leads to the congestion window being increased while packet loss (or an increase in delay as in the Vegas algorithm [1] TODO: How much detail is necessary in the related work section?) lead to the congestion window being lowered according to a set of fixed rules. Algorithms such as Cubic and Compound [6, 10] improve congestion control specifically with respect to so-called long fat networks with a high round-trip time and bandwidth. Proportional rate reduction [4] aims to make the reduction of the rate that occurs in case of congestion more smooth and steady in time to avoid the bursty behavior that occurred upon loss in previous TCP congestion control variants. On the other hand, BBR [2] does not use a set of fixed rules such as previous congestion control algorithms but instead estimates a model of the network path by using measurements of RTT and throughput and then adjusts its sending rate so that it uses the maximum bandwidth according to its network model while trying not to fill up queues. A similar approach is PCC [3], which also uses measurements to find an optimum sending rate with respect to a defined utility function (e.g. maximizing throughput while minimizing packet loss). However,
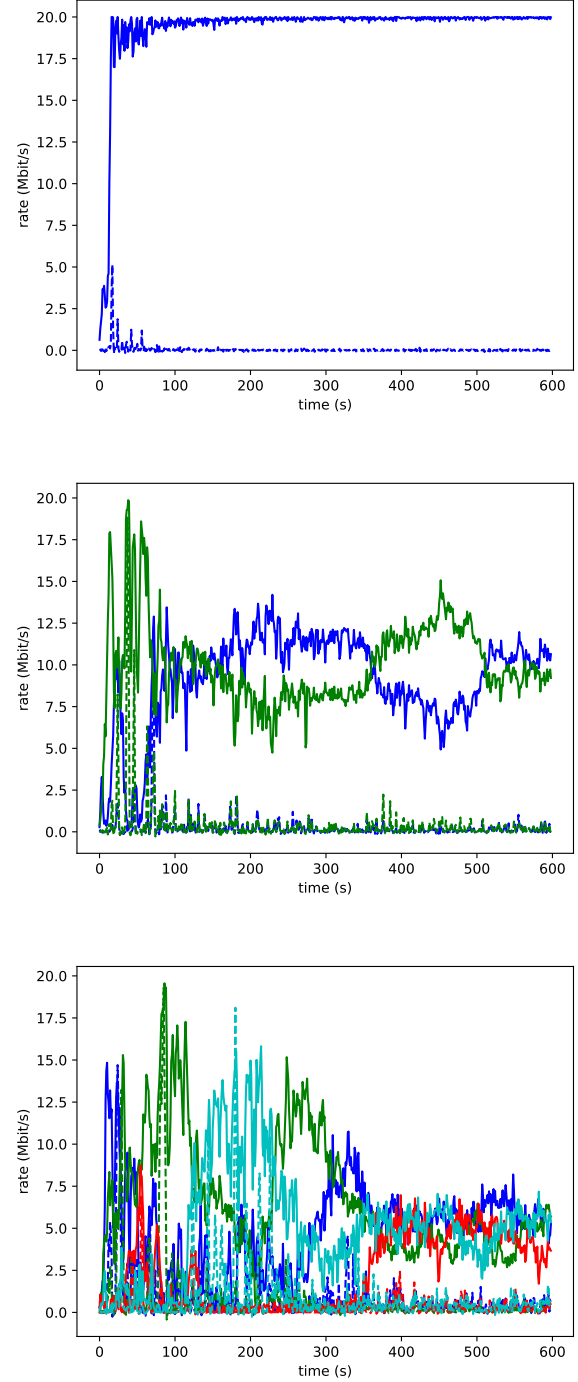

Figure 1: TODO: figures don't look that appealing do they? The throughput (solid line) and the loss rate (dashed line) when sharing a bottleneck link of 20 Mbit/s with a two-way end-to-end delay of 50 ms and a very small buffer of $\frac{1}{10}$ bandwidth delay product between 1, 2 and 4 senders (from top to bottom). One can see that the time to convergence increases as the number of senders increases, which is caused by the fact that the more senders use the link, the fewer packets each sender receives per unit of time for training. Also, the complexity of the learning problem increases with more senders as the environment becomes more dynamic and unpredictable.

PCC doesn't maintain a congestion window but instead modulates the sending rate.

Besides the aforementioned human-designed algorithms there have been attempts to use machine learning to improve congestion control. Geurts et al. [5] train a classifier off-line to determine whether packet loss is caused by congestion or by the link layer. If it is caused by the link layer, TCP does not alter the window. Otherwise – if a packet was lost due to congestion according to the classifier – it uses a traditional TCP congestion control algorithm. Winstein and Balakrishnan [11] train a machine learning solution called *Remy* that finds an optimum congestion control for a given range of network parameters. For instance, one could find an optimum congestion control for networks with an RTT of 50-100 ms and link speeds of 100-500 Mbit/s. After a lengthy training procedure Remy finds an optimum congestion control algorithm for the specified networks on a per acknowledgement basis (updating the window upon receiving an acknowledgement alike traditional congestion control algorithms).

Our goal is to find an optimum congestion control algorithm on a per acknowledgement basis similar to Remy that, however, can be trained on-line. Such a solution could be used in a purely on-line fashion, in a off-line fashion like Remy or a combination of both: One could pre-train a generic congestion control algorithm that works reasonably well for every network that then gets refined during on-line training according to the current network circumstances.

To this end we propose Partial Action Learning which is based on the Asynchronous Advantage Actor Critic framework [9], which has been demonstrated to be able to learn to play a wide range of video games and commonly outperform human players. In particular, it is a good choice for congestion control as it has been proven to deliver good performance and high training speed for playing video games and also for networking related tasks such as Adaptive Bitrate Algorithms [8] TODO: This motivation doesn't sound convincing yet, does it?. Furthermore, as we will show, it is conceptually possible to use it for on-line training although – to our knowledge – this has not been done until now.

2.2 Actor Critic Learning

The *Partial Action Learning* (PAL) (see 3.1) framework is based on the Actor Critic framework for neural networks proposed by Mnih et al. [9], which we outline (with a focus on intuition over rigor) in this section.

There are two neural networks, the Actor Network and the Critic Network. Given a state, the Actor Network outputs what it deems to be the optimum action to perform in that certain state. The Critic Network estimates what long-term reward can be expected in this state. So an action is considered good if it achieved a long-term reward that is higher than the long-term reward expected by the Critic Network and it is considered bad if the reward was lower than expected. The long-term reward is implemented as an exponentially weighted moving average of future rewards. So if a high reward can be achieved right now this is more favorable than if it can be achieved in the future. However it can also be beneficial to get a low reward now and instead get a very large one in the future.

2.2.1 Critic Network. The Critic Network outputs the expected long-term reward $V(s_t; \theta_v)$ given a state s_t at time step t and the parameters (neural network weights) of the critic network θ_v.

With r_t being the reward that was received at time t and γ being a factor with $0 < \gamma \leq 1$, which stands for the influence that future reward has on the moving average, we define the expected long-term reward at time step t as

$$R_t = \gamma r_t + (1 - \gamma)R_{t+1},$$

One can see that to compute R_t at time step t one has to look infinitely far into the future to get all future rewards needed to compute the exponentially weighted moving average at time step t. Thus, one usually only carries out t_{max} steps (for example 20) and then uses an estimation of $R_{t_{max}+1}$ provided by the critic network as the continuation TODO: Clearer than before?. The only exception from this is when the task to carry out is over before reaching t_{max} (e.g. the TCP flow is over). In this case, one can either use a prediction from the critic network as a continuation as usually or simply let the moving average end without continuation (ignoring the term weighted with $(\gamma - 1)$ at the last step).

The loss function (not in the sense of packet loss but in the sense of the loss of a machine learning problem), which the critic network tries to minimize at each time step, is the square of the difference of the actual long-term reward received and the expected long-term reward

$$l_{v,t} = (R_t - V(s_t; \theta))^2.$$

2.2.2 Actor Network. The Actor Network outputs a probability distribution from which the action a_t at time step t is randomly sampled. We use the mean μ and the standard deviation σ to parametrize a normal distribution. The main idea is that the network learns to output the right mean at the right time step to maximize the future reward and that it uses the standard deviation to try out new actions, which could yield a better than expected reward.

Each time we take an action a_t we also let the critic network make an estimation $V(s_t; \theta)$, which we abbreviate as v_t, given the current neural network weights θ and the current state s_t. The loss function of the actor network can now be described as this: How likely was it (on a logarithmic scale) to take action a_t given the probability distribution at time t and how beneficial was this action when comparing to the critic's estimation? For example, if an action was very successful and highly unlikely, the actor network's are updated to make actions like this more likely in the future.

The actor also aims at increasing the entropy of the probability density function ($\frac{1}{2} \log(2\pi e \sigma^2)$) to encourage exploration: Otherwise the actor could always output similar actions but never actually take the action that yields the best performance. We designate the entropy as H and introduce a parameter β which specifies magnitude of the influence of the entropy.

$$l_{a,t} = -\log(\pi(a_t \mid s_t; \theta))(R_t - v_t) - \beta H(\pi(s_t; \theta)).$$

3 METHOD

3.1 Partial Action Learning

The key difference between PAL and previous approaches to RL is that in classical RL an action is always followed by a reward and a
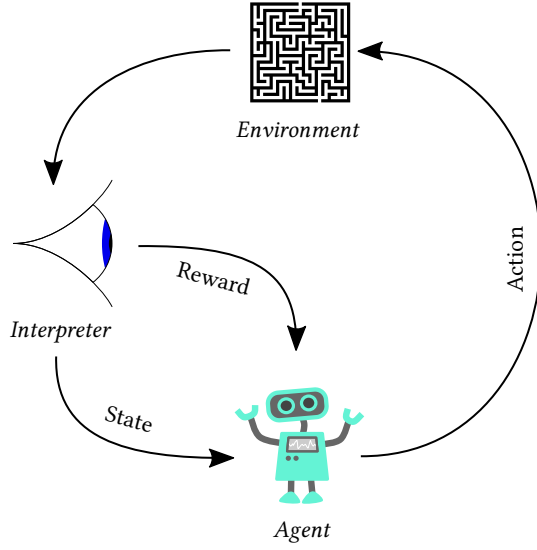
Figure 2: The classical reinforcement learning approach.[a]

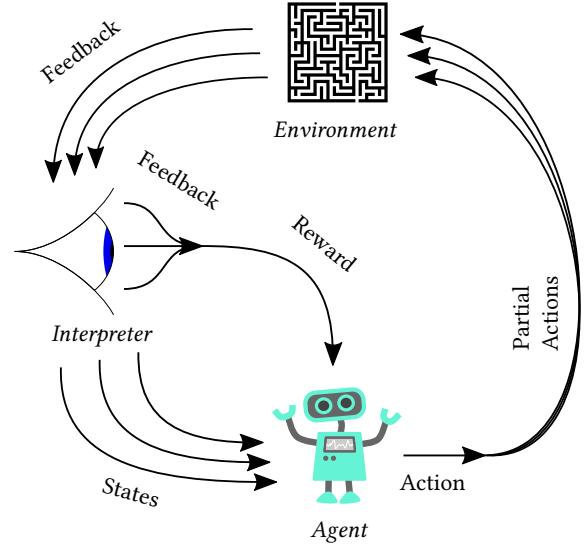[a]adapted from https://upload.wikimedia.org/wikipedia/commons/1/1b/ Reinforcement_learning_diagram.svg

Figure 3: Partial Action Learning: An action consists of zero or more partial actions which trigger feedback upon interacting with the environment. Each feedback updates the state. The value and actor networks are updated upon receiving all feedback of one action.[a]

[a]adapted from https://upload.wikimedia.org/wikipedia/commons/1/1b/ Reinforcement_learning_diagram.svg

reward is always followed by an action. In our proposed concept, however, it is possible to take new actions while previous actions haven't received their rewards yet.

Another major difference in PAL is that one action generates a number of partial actions (≥ 0) (see Figure 3). Each partial action generates feedback upon interacting with the environment. Upon receiving feedback for a partial action, the agent determines the current state and triggers a new action. When all feedbacks of one action were received, the agent combines them to form the reward and updates the critic and actor networks.

In Algorithm 1 (in the appendix) we show the code that runs in each of the agents (in the congestion control scenario, one agent corresponds to a sender). It is possible to have several agents which share a set of neural weights (which is usually done when performing off-line training [8, 9]) but one can also use separate weights for each agent, which is more realistic in case of congestion control in the Internet, as it is not sensible that different senders share a set of neural network weights over the Internet.

3.2 Congestion Control Specifics

The motivation for PAL is that classical reinforcement learning assumes that a reward follows an action and vice-versa (see Figure 2). TODO: Too much repetition of this? However, in the case of congestion control, it is desirable to perform a new action without having received a reward for the previous action. For example, imagine that we receive an acknowledgement. Now we have to sample an action and wait for the reward, and it takes one RTT until the first acknowledgements (which make up the reward) are received. If we now receive another acknowledgement before receiving the reward of the previous action, classical RL cannot handle this situation because of the asynchronicity of actions and rewards. Thus the

Asynchronous Actor Critic framework as described by [9] cannot be applied to congestion control as it assumes that actions and rewards are synchronized and so we have to use Partial Action Learning (see subsection 3.1). We describe the overall workings of PAL for congestion control using a petri net (see Figure 4).

To use PAL for congestion control we first have to define the correct semantics for this specific use case and we have to explicitly state how the state, reward etc. are defined. Furthermore, we have to define how the actor and critic network explicitly work in case of congestion control. In the following a time step t corresponds to the reception of an acknowledgement. The beginning of the flow, before any packet is sent, corresponds to time step 0.

s_t The state describes the current "congestion state". The following features are included in it:
- the time between the last two packets that were sent
- the time between the last two packets that were received
- the RTT of the last received packet
- whether the last packet was lost where 1 signifies that the last packet was lost while 0 indicates that it was received correctly
- the current congestion window as a real number

Of each of these features the feature itself is included as well as one exponentially weighted moving average (ewma) with a factor α of $\frac{1}{8}$ and one with an α of $\frac{1}{256}$ (the concept of using ewmas with these specific values of α was first introduced in *Remy* [11]). Each time an acknowledgement
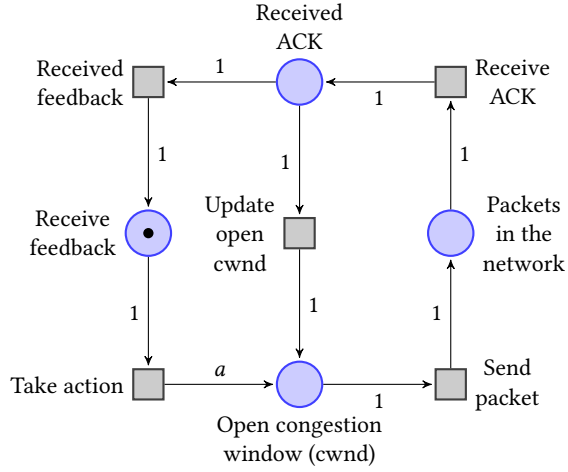
Figure 4: A petri net describing the congestion control mechanism. We start with one token in the state *Received feedback,* which means that we can take an action right in the beginning. The concept is the following: If there is at least one token in the *Open congestion window* send a packet (for the sake of simplicity in this example we assume that the congestion window is in packets and not bytes). The *Open congestion window* signifies the congestion window minus the packets that are currently unacknowledged. The token now goes to the network and after some time the acknowledgement (ACK) (or negative-acknowledgement (NACK) or timeout if the packet gets lost) for the packet is received in the *Received ACK* state. From here one token goes to the *Open congestion window* meaning that when we get an ACK, NACK (or timeout) we can send another packet. Furthermore the state *Received feedback* receives a token upon receiving an ACK/NACK/timeout. *Take action* adds *a* tokens to the *Open congestion window,* where *a* is a real number (possibly also negative); thus in each state there can also be a real number of tokens (e.g. 2.34 tokens are possible in the *Open congestion window*). However, we define that the congestion window can never be smaller than 1, which means that the sum of all tokens in the *Open congestion window., Received ACK* and *Packets in the network* states can never be smaller than 1.

is received, the state is updated and the actor network is asked for the next action.

a_t Based on a given state and the history of previous states (because we use Gated Recurrent Units (GRUs) and thus can also consider previous states), the actor network returns an action a_t, which is a real number that stands for the change of the congestion window.

r_t The reward is a tuple of five reward metrics. For each reward metric there is also an output of the critic network that predicts the expected long-term average of this reward metric given the current state.
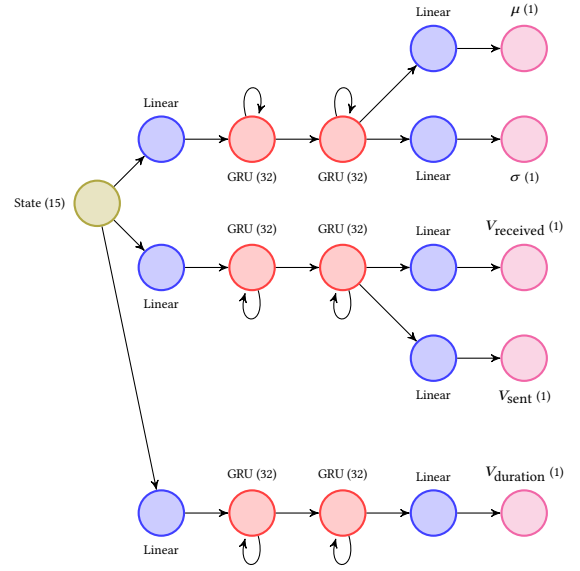

Figure 5: An overview of the complete neural network being used. Ocher stands for the input, blue for linear layers, red for GRU layers and purple for the outputs of the neural network. The numbers in parentheses next to each label stand for the width of that layer (i.e. the number of neurons except for the inputs and outputs).

v_t The value is a tuple of the expected average reward estimated by the critic network (see subsubsection 3.2.1) for each of the reward metrics.

We use the following reward metrics:

$r_{\text{sent}, t}$ is the sum of all bytes in all packets the sender sent during time step t.

$r_{\text{received}, t}$ is the sum of all bytes in all packets that the sender sent during time step t and that were not lost (so they were acknowledged at some point by the receiver).

$r_{\text{duration}, t}$ is the time between receiving the last packet and receiving this packet ("inter acknowledgement reception time") summed over all packets that the sender sent and that were not lost. Example: The sender sent three packets, of which two were not lost. The time between receiving the ACK of the first packet that wasn't lost and the previous one is 5 ms and the time between receiving the ACK of the second packet that wasn't lost and the previous ACK is 5 ms. So for this time step $r_{\text{duration}, t}$ is 11 ms.

The overall structure of the neural network for both the value and the actor network is depicted in Figure 5. Having these reward metrics one can compose a variety of functions with their long-term averages. How these reward metrics can be used in reinforcement learning is described in subsection 3.3.

3.2.1 Critic Network. In the case of congestion control, the loss function $l_{\text{v}, t}$ of the critic network is actually the sum of the squares of the difference for each of the expected long-term averages and

the empirically found averages for each reward metric:

$$l_{v,t} = \left(R_{\text{received},t} - V_{\text{received}}(s_t;\theta_v)\right)^2$$
$$+ \left(R_{\text{sent},t} - V_{\text{sent}}(s_t;\theta_v)\right)^2$$
$$+ \left(R_{\text{duration},t} - V_{\text{duration}}(s_t;\theta_v)\right)^2$$

3.2.2 Actor Network. Each time an action a_t is requested, a change to the congestion window is sampled from the current normal distribution defined by the parameters μ and σ. However, we define that the window can never fall below 1: $w_{t+1} = \max(w_t + a_t, 1)$. At the beginning of a flow, the window starts as 1 too.

The actor network minimizes the loss function

$$l_{a,t} = -\log\left(\pi\left(a_t \mid s_t;\theta_a\right)\right)$$
$$\left(U_{\text{measured},t} - U_{\text{expected},t}\right)$$
$$- \beta H\left(\pi\left(s_t;\theta_a\right)\right)$$

where U can be any utility function defined based on some of the (long term averages of the) reward metrics. Thus, in simple terms, the actor network considers if an action improved the actual experienced utility was compared to the expected one and adjusts the neural network accordingly.

3.3 Utility function

We use a reward function that is similar to the one used by PCC [3].

$$U_t = \text{throughput} - \alpha \times \text{lost throughput}$$

α is a parameter that determines how strongly packet loss is discouraged. We can define this utility function as follows using the previously defined reward metrics:

$$U_t = \frac{R_{\text{received},t}}{R_{\text{duration},t}} - \alpha\frac{R_{\text{sent},t} - R_{\text{received},t}}{R_{\text{duration},t}}$$

The equivalent utility is also defined using the estimates of the value networks and it can be determined whether an action achieved a performance that was higher or lower than expected.

3.4 Example

In this section, an example of the overall procedure is outlined.

(1) Receive an ACK
(2) Update the internal state to reflect the information that this ACK provides (for example, the state consists of the round trip time experienced by this packet, the time since we receive the last ACK etc.)
(3) Sample an increase for the congestion window from the Actor Network (e.g. 0.24) and add it to the current congestion window.
(4) (a) If this was the last ACK, which completes a previous action (e.g. a previous action resulted in 3 packets being sent; so if this is the ACK for the third (and last) packet of this previous action, we got all the ACKs to compute the reward), combine all partial rewards to form the actual reward for that action.

(b) If we now have $t_{\max}$ rewards (set to 20) then we also update the neural network.

4 RESULTS

5 DISCUSSION

5.1 Best practices for on-line reinforcement learning

6 CONCLUSION

APPENDIX

Algorithm 1 Partial Action Learning – pseudocode for each agent. It is possible that the agents share the global weights θ_g, which is reasonable when learning off-line. Otherwise an individual copy of them is kept by each agent. By default, all weights are initialized randomly in the beginning but it is also possible to use a pre-trained neural network for initialization. As we use Gated Recurrent Units (GRUs) their output is stored in h and fed into them at the next time step.

1: **loop**
2: $\quad l_{\text{actions}} \leftarrow []$
3: $\quad l_{\text{states}} \leftarrow []$
4: $\quad l_{\text{values}} \leftarrow []$
5: $\quad l_{\text{rewards}} \leftarrow []$
6: $\quad l_{\text{values'}} \leftarrow []$
7: $\quad l_{\text{snapshots}} \leftarrow []$
8: $\quad l_{\text{hiddenStates}} \leftarrow []$
9: $\quad t \leftarrow 0$
10: $\quad s_0 \leftarrow \text{initialState}()$
11: $\quad h \leftarrow \text{zeros}()$
12: $\quad \theta \leftarrow \theta_g$
13: $\quad$ **repeat**
14: $\quad\quad l_{\text{values'}}.\text{append}(V(s_t; \theta))$
15: $\quad\quad$ **if** UPDATEWEIGHTS **then**
16: $\quad\quad\quad \theta_a \leftarrow \theta_g$
17: $\quad\quad\quad l_{\text{snapshots}}.\text{append}(\theta)$
18: $\quad\quad\quad l_{\text{hiddenStates}}.\text{append}(h)$
19: $\quad\quad$ **end if**
20: $\quad\quad l_{\text{states}}.\text{append}(s_t)$
21: $\quad\quad$ Sample a_t from the
$\quad\quad\quad$ Actor Network's probability distribution
22: $\quad\quad l_{\text{actions}}.\text{append}(a_t)$
23: $\quad\quad l_{\text{values}}.\text{append}(V(s_t; \theta_v))$ ▷ Possibly different from the
$\quad\quad\quad$ value in $l_{\text{values'}}$ as the weights could have
$\quad\quad\quad$ been updated in the meantime.
24: $\quad\quad l_{\text{partialActions}, t} \leftarrow \text{partialActions}(a_t)$
25: $\quad\quad$ **for all** partial actions $a_{\text{p}, i, t}$ in $l_{\text{partialActions}, t}$ **do**
26: $\quad\quad\quad$ Take partial action $a_{\text{p}, i, t}$
27: $\quad\quad$ **end for**
28: $\quad\quad$ Wait for the next feedback $r_{\text{p}, j, t'}$
$\quad\quad\quad$ where $t' \leq t$ and $0 \leq j \leq \#(l_{\text{partialActions}, t'})$
29: $\quad\quad$ **if** all feedback of $a_{t'}$ was received **then**
30: $\quad\quad\quad r_{t'} \leftarrow$ reward w.r.t all feedback $r_{\text{p}, k, t'}$
$\quad\quad\quad\quad$ where $0 \leq k \leq \#(l_{\text{partialActions}, t'})$
31: $\quad\quad\quad l_{\text{rewards}}.\text{append}(r_{t'})$
32: $\quad\quad\quad$ **if** ENOUGHREWARDS or the episode is over **then**
33: $\quad\quad\quad\quad$ COMPUTEGRADIENTS
34: $\quad\quad\quad$ **end if**
35: $\quad\quad$ **end if**
36: $\quad\quad$ Generate s_{t+1} using $r_{\text{p}_{t'}}$
37: $\quad\quad t \leftarrow t + 1$
38: $\quad$ **until** reaching the end the episode
39: **end loop**

Algorithm 2 Partial Action Learning – procedure which computes and applies the gradients.

1: **function** COMPUTEGRADIENTS
2:    $t_{\text{end}} = \#(l_{\text{rewards}})$
3:    $\theta_{\text{backup}} \leftarrow \theta$
4:    $h_{\text{backup}} \leftarrow h$
5:    $\theta \leftarrow l_{\text{snapshots}}[0]$
6:    $h \leftarrow l_{\text{hiddenStates}}[0]$
7:    $R_{i+1} \leftarrow$ last element of $l_{\text{values'}}$
8:    **for** $i \leftarrow t_{\text{end}} - 1..0$ **do**
9:       $R_i \leftarrow (r_i + \gamma R_{i+1})(1 - \gamma)$
10:      $a \leftarrow l_{\text{actions}}[i]$
11:      $s \leftarrow l_{\text{states}}[i]$
12:      $v \leftarrow l_{\text{values}}[i]$
13:      $d\theta \leftarrow -\dfrac{\partial \log(\pi(a \mid s;\theta_{\text{a}}))(R_i - v)}{\partial \theta}$
$\qquad\qquad - \dfrac{\partial \beta H(\pi(s;\theta_a))}{\partial \theta}$
$\qquad\qquad + \dfrac{\partial (R_i - v)}{\partial \theta}$
14:      $\theta_{\text{g}} \leftarrow \theta_{\text{g}} + d\theta$
15:    **end for**
16:    Remove first t_{end} elements from
     $l_{\text{actions}}, l_{\text{states}}, l_{\text{values}}, l_{\text{rewards}}, l_{\text{values'}}$
17:    Remove the first element from $l_{\text{snapshots}}$
18:    Remove the first element from $l_{\text{hiddenStates}}$
19:    $\theta = \theta_{\text{backup}}$
20:    $h = h_{\text{backup}}$
21: **end function**

Algorithm 3 Procedure which determines when to synchronize the neural network weights from the global weights in case of congestion control. We assume that the window at time step t is stored in a list l_{windows} and that the number of bytes sent in each time step is stored in l_{sent}. Alternatively to this procedure one can simply synchronize the weights after each t_{max} actions where t_{max} is an arbitrarily chosen integer greater than 0.

1: **function** UPDATEWEIGHTS
2:    $i \leftarrow 0$
3:    **while** $i < \#(l_{\text{actions}})$ **do**
4:      $w \leftarrow \lfloor l_{\text{windows}}[i] + l_{\text{actions}}[i] \rfloor$
5:      **for** $j \leftarrow i..\#(l_{\text{actions}}) - 1$ **do**
6:        $w \leftarrow w - l_{\text{sent}}[j]$
7:        **if** $w \leq 0$ **then**
8:          $i \leftarrow j$
9:          **break**
10:        **end if**
11:      **end for**
12:      **if** $i = \#(l_{\text{windows}}) - 1$ **then**
13:        **return true**
14:      **end if**
15:    **end while**
16:    **return false**
17: **end function**

Algorithm 4 Procedure which determines when to start computing the gradients in case of congestion control. We assume that the window at time step t is stored in a list l_{windows} and that the number of bytes sent in each time step is stored in l_{sent}. Alternatively to this procedure one can simply compute the gradients when t_{max} rewards have been received where t_{max} is an arbitrarily chosen integer greater than 0.

1: **function** ENOUGHREWARDS
2:    w $\leftarrow \lfloor l_{\text{windows}}[0] + l_{\text{actions}}[0] \rfloor$
3:    **for** $i \leftarrow 0..\#(l_{\text{rewards}}) - 1$ **do**
4:      w $\leftarrow$ w $- l_{\text{sent}}[i]$
5:      **if** w ≤ 0 **then**
6:        **return true**
7:      **end if**
8:    **end for**
9:    **return false**
10: **end function**

REFERENCES

[1] Lawrence S. Brakmo and Larry L. Peterson. 1995. TCP Vegas: End to end congestion avoidance on a global Internet. *IEEE Journal on selected Areas in communications* 13, 8 (1995), 1465–1480.
[2] Neal Cardwell, Yuchung Cheng, C Stephen Gunn, Soheil Hassas Yeganeh, and Van Jacobson. 2016. BBR: Congestion-based congestion control. *Queue* 14, 5 (2016), 50.
[3] Mo Dong, Qingxi Li, Doron Zarchy, Philip Brighten Godfrey, and Michael Schapira. 2015. PCC: Re-architecting Congestion Control for Consistent High Performance.. In *NSDI*. 395–408.
[4] Nandita Dukkipati, Matt Mathis, Yuchung Cheng, and Monia Ghobadi. 2011. Proportional rate reduction for TCP. In *Proceedings of the 2011 ACM SIGCOMM conference on Internet measurement conference*. ACM, 155–170.
[5] Pierre Geurts, Ibtissam El Khayat, and Guy Leduc. 2004. A machine learning approach to improve congestion control over wireless computer networks. In *Data Mining, 2004. ICDM'04. Fourth IEEE International Conference on*. IEEE, 383–386.
[6] Sangtae Ha, Injong Rhee, and Lisong Xu. 2008. CUBIC: a new TCP-friendly high-speed TCP variant. *ACM SIGOPS Operating Systems Review* 42, 5 (2008), 64–74.
[7] Van Jacobson. 1988. Congestion avoidance and control. In *ACM SIGCOMM computer communication review*, Vol. 18. ACM, 314–329.
[8] Hongzi Mao, Ravi Netravali, and Mohammad Alizadeh. 2017. Neural Adaptive Video Streaming with Pensieve. In *Proceedings of the Conference of the ACM Special Interest Group on Data Communication (SIGCOMM '17)*. ACM, New York, NY, USA, 197–210. DOI:http://dx.doi.org/10.1145/3098822.3098843
[9] Volodymyr Mnih, Adria Puigdomenech Badia, Mehdi Mirza, Alex Graves, Timothy Lillicrap, Tim Harley, David Silver, and Koray Kavukcuoglu. 2016. Asynchronous methods for deep reinforcement learning. In *International Conference on Machine Learning*. 1928–1937.
[10] Kun Tan, Jingmin Song, Qian Zhang, and Murad Sridharan. 2006. A compound TCP approach for high-speed and long distance networks. In *Proceedings-IEEE INFOCOM*.
[11] Keith Winstein and Hari Balakrishnan. 2013. Tcp ex machina: Computer-generated congestion control. In *ACM SIGCOMM Computer Communication Review*, Vol. 43. ACM, 123–134.