Deep Reinforcement Learning in Real-Time Bidding

Author:
Oskar Stigland
Bachelor Thesis
Fall 2018

Supervisors:
Alexandros Sopasakis
Morten Arngren
Vlad Sandulescu

LUND UNIVERSITY

Centre for Mathematical Sciences
Numerical Analysis

Abstract

This segment should describe the contents of the thesis, how the experiments etc have been carried out, what type of methods have been used and a short summary of the results of the project. Include GitHub link to repository.

Acknowledgements

I want to thank...

Contents

Chapter 1

Reinforcement Learning

The fundamental purpose of reinforcement learning is to design agents with the ability to successfully navigate through environments from which they have no prior experience. This does not necessarily mean that they have no prior knowledge of the environment whatsoever, although this is often the case as we shall see later, but rather that they haven't taken any actions in the environment previously; they have no idea of what kind of consequences or rewards follow from different actions.

Imagine a kid trying to learn how to ride a bike. The kid might understand how a bike works, e.g. that turning the handlebars to the right makes the bike turn right and that pushing the bike pedals makes the bike accelerate and go forward, and so on. However, there are a few things that only experience can teach. For example, it's difficult, if not impossible, to understand beforehand just how much the handlebar will make the front wheel turn. Similarly, it's hard to understand how much the bike will accelerate if we push the bike pedals forward or how harshly it will brake if we push the pedals backwards. Most importantly, it's impossible to know how much it will actually hurt to hit the ground, or if it will even hurt, when you fall off the bike if you haven't already done it, or how exhilarating it is to bike fast.

The latter example is of importance for reinforcement learning, since consequence and reward are how we make sure that an agent learns to behave in an optimal way in some environment. Just like the kid experiences pain and failure when it falls off the bike, we make sure our agent receives negative or low numerical rewards when choosing "bad" actions and, conversely, that it receives positive numerical rewards when it acts in a "good" way.

1.1 Markov Decision Processes

The most common way to model a reinforcement-learning problem is through the Markov Decision Process (MDP). In this thesis, and in reinforcement learning in general, the finite MDP is of most importance, where there is a finite number of combinations of situations (or *states*) and actions as well as a finite (discretized) interval of rewards. In defining the general framework of a finite MDP, and the notation to be used later in this thesis, I will follow Sutton and Barto (2018).

We consider a finite series of time steps, $t = 1, 2, \ldots, T$, where T is the time of termination for whatever we're doing, and denote the action, state and reward at time t by A_t, S_t and R_t, respectively. We define a finite set for each: $A_t \in \mathcal{A}$, $S_t \in \mathcal{S}$ and $R_t \in \mathcal{R} \subset \mathbb{R}$. First, we want to consider the joint probability of some state, s', and some reward, r, following the choice of a certain action, a, in a certain state, s:

$$p(s', r, s, a) \triangleq P(S_{t+1} = s', R_{t+1} = r | S_t = s, A_t = a)$$

$\forall\, s', s \in \mathcal{S}$, $\forall\, r \in \mathcal{R}$ and $\forall\, a \in \mathcal{A}$. We have that $p : \mathcal{S} \times \mathcal{S} \times \mathcal{A} \times \mathcal{R} \to [0, 1]$. Using this notation, we define value of taking a certain action, a, in a certain state, s, as

$$q(s, a) \triangleq \mathbb{E}\left[R_{t+1} | S_t = s, A_t = a\right] = \sum_{r \in \mathcal{R}} r \sum_{s' \in \mathcal{S}} p(s', r, s, a)$$

such that $q : \mathcal{S} \times \mathcal{A} \to \mathbb{R}$. Considering our biking kid, we might have a state in which the bike is on the crest of a hill and is just about to start rolling downwards. Our function, $q(s, a)$, then maps different actions, e.g. accelerating and breaking, to their perceived value. When finding these values, there's an important aspect to consider. For example, acceleration might yield some short-term exhilaration, but it also means some future risk as the kid will have less control over the bike.

Making a choice isn't just about weighing different immediate rewards against each other, it's also about weighing the present against the future, balancing the short-term and the long-term. This is also true for a reinforcement-learning agent, which we formalize using a *discount factor*, $0 \leq \gamma \leq 1$. A low γ means that the agent is *myopic*, prioritizing short-term rewards, while a high γ means that the agent will give consideration to future rewards.

Using the discount factor, we formulate the expected return at time t as

$$G_t \triangleq R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k$$

which gives us a recursive relationship, since

$$G_t = R_{t+1} + \gamma R_{t+2} + \gamma^2 R_{t+3} + \cdots + \gamma^{T-t-1} R_T$$

$$= R_{t+1} + \gamma(R_{t+2} + \gamma R_{t+3} + \cdots + \gamma^{T-t-2} R_t) = R_{t+1} + \gamma G_{t+1}$$

such that $G_t = R_{t+1} + \gamma G_{t+1}$. We use this recursive relationship to re-define the function $q(s, a)$:

$$\vdots$$

$$\vdots$$

for all $s \in \mathcal{S}$ and $a \in \mathcal{A}$. The reason for this re-definition will become clear shortly. The function $q(s, a)$ is called the *action-value function* and will be one of the most important conceptual features of this thesis.

1.1.1 An Optimal Policy

In reinforcement learning, a policy can be strictly defined as a mapping from states, s, to probabilities of selecting certain actions, a, in those states. A policy is usually denoted by π and we can hence express it as

$$\pi(a, s) = P(A_t = a | S_t = s)$$

$\forall\, a \in \mathcal{A}$, $\forall\, s \in \mathcal{S}$, and for $t = 1, 2, \ldots, T$. If an agent follows a policy π, we say that $q_\pi(s, a)$ is the *action-value function for policy* π. If a particular policy π_* has the property that $q_{\pi_*}(s, a) \geq q_\pi(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ and for all other policies π, we call it the *optimal policy.* For this policy, we denote the action-value function by $q_*(s, a)$ and define it as

$$q_*(s, a) \triangleq \max_\pi q_\pi(s, a), \quad \forall\, a \in \mathcal{A}, \forall\, s \in \mathcal{S}$$

Hence, the purpose of the optimal policy is to maximize the expected return, G_t, at any time $t = 0, 1, 2, \ldots, T$. We define $R = G_0$, i.e. such that the expected return for a whole period is denoted by R.

1.1.2 The Bellman Optimality Equation

We consider the definition of $q(s, a)$ above. Assuming that we are following the optimal policy, π_*, we can rewrite the action-value function as a recursive relationship:

$$q_*(s, a) = \mathbb{E}\left[R_{t+1} + \gamma \max_{a'} q_*(S_{t+1}, a') \middle| S_t = s, A_t = a \right]$$

$$= \sum_{s',r} p(s', r, s, a) \left[r + \gamma \max_{a'} q_*(s', a') \right]$$

This equation says that the expected return from taking an action a in a state s when following the optimal policy corresponds to the immediate expected reward and the expected return from following the optimal policy in the next state. This is known as the Bellman optimality equation for the action-value function. While this is analytically nice, it's applicability is constrained by computational complexity. Hence, many reinforcement-learning techniques, such as Monte Carlo methods, aim to approximate $q_*(s, a)$. This is also true for the method which will later be introduced as the *Deep Q-Network*.

1.1.3 Constrained Markov Decision Processes

So far, we've been concerned with an agent who's making decisions to maximize a single metric: the expected reward. In this case, we want to choose a policy π such that

$$\pi = \arg\max_{\pi} \mathbb{E}[R|\pi] = \arg\max_{\pi} \left[\sum_{t=1}^{T} \gamma^{t-1} R_t \middle| \pi \right]$$

which we've previously defined as the optimal policy, π_*. However, we've only been concerned with unconstrained maximization. It is not clear that π_* is an optimal policy when we impose constraints on the agents. We refer to such a case as a *Constrained Markov Decision Process* (CMDP). We define $C = \sum_{t=1}^{T} \gamma^{t-1} C_t$, where C_t is the cost at time t, and instead consider the problem of finding π such that

$$\begin{aligned} \max_{\pi} \quad & \mathbb{E}\left[R|\pi\right] \\ \text{s.t.} \quad & \mathbb{E}\left[C|\pi\right] \leq c \end{aligned}$$

where c is our cost constraint. How do we make this fit into the reinforcement-learning framework? Geibel (2007) discusses a number of methods fitted to different CMDP problems, one of which is to expand the state space, $\mathcal{S}$, to include the cost constraint. Instead of just considering the normal state-relevant parameters when taking an action, we also consider the cost incurred and if the constraint has been reached the agent will either be incapacitated or the period will be terminated. This is the approach that will be followed in this thesis. **Explanation why?**

1.2 Exploration and exploitation

As mentioned in the introduction to this chapter, reinforcement-learning techniques aim to deploy agents into new environments. This means that they have to *explore* the environment before being able to act intelligently, essentially taking random actions to see what happens. When the agent is acting intelligently and taking the decisions that it knows maximizes the expected return, we say that it is *exploiting*. When the agent is only exploiting and not exploring, we call it *greedy*. Hence, the optimal policy, as defined by $q_*(s, a)$, is greedy since we're always choosing the actions that will maximize the expected return.

The exploration-exploitation trade-off is one of the most important aspects of reinforcement learning. On the one hand we want the agent to be as well-informed as possible about the actions it's taking, but on the other hand we want it to gain as much reward as possible; more exploration means less exploitation, and vice versa. This is usually solved by a so-called ϵ-greedy policy, which means that the agents exploits with a probability $1-\epsilon$ and explores with a probability ϵ, where $0 \leq \epsilon \leq 1$.

When the agent is learning, we want it to explore as much as possible, i.e. to have a high ϵ. Conversely, when the agent has finished learning, we want it to exploit, i.e. having a low ϵ or even $\epsilon = 0$. In practice, this can be solved by a number of ways. Often, it is the case that the agent start out with $\epsilon \geq 0.9$ and then lets ϵ decay over time, e.g. linearly or exponentially, according to some fixed rate. In this thesis, several approaches will be tried later on, but focus will be on a ϵ-greedy policy with an exponentially decaying ϵ. **Discussion on why?** The figure below illustrates an agent following an ϵ-greedy policy.

1.3 Model-free and model...

...

1.4 Q-Learning

One of the most famous RL methods is the model-free *Q-learning algorithm*, which, as the name suggests, is concerned with directly estimating the action-value function. While Q-learning deserves a longer theoretical background and discussion of its convergence properties, this will make for a shorter introduction as we will ultimately not be concerned with the Q-learning algorithm, but rather with a variant of Q-learning for which we can't necessarily make claims about stability and convergence.

We start by initializing some arbitrary action-value function $Q_0(s, a)$ for all $s \in \mathcal{S}, a \in \mathcal{A}$ and as the agent is exploring the environment (as well as when it's exploiting), we're continually making incremental updates for the action-value function:

$$Q_{n+1}(S_t, A_t) = Q_n(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q_n(S_{t+1}, a) - Q_n(S_t, A_t) \right]$$

where α is the learning rate. That is, when observing S_t, A_t and R_{t+1}, we update the action-value function by the scaled difference between the old value, $Q_n(S_t, A_t)$, and the sum of the immediate reward and the discounted expected return for the next state under a greedy policy, $R_{t+1} + \gamma \max_a Q_n(S_{t+1}, a)$. If we look closely, this is actually a familiar sight. Setting $\alpha = 1$, we get

$$\begin{aligned} Q_{n+1}(S_t, A_t) &= Q_n(S_t, A_t) + \left[R_{t+1} + \gamma \max_a Q_n(S_{t+1}, a) - Q_n(S_t, A_t) \right] \\ &= R_{t+1} + \gamma \max_a Q_n(S_{t+1}, a) \end{aligned}$$

which is analogous to the Bellman optimality equation for the action-value function for a greedy policy, i.e. where we choose the best action with probability 1. This is the update rule we will be using when constructing our agent later, with $\alpha = 0.001$. It's important to note that the Q-learning update occurs at almost every step, meaning that the algorithm keeps updating the values for all state-action pairs even as ϵ decreases and the agent exits the exploring phase.

One of the problems with Q-learning is that we might run into trouble in large-scale systems since it's hard to visit every state-action pair a sufficient amount of times to get a good estimate of their values, especially when we have to balance exploration and exploitation. This is why we are now turning to the next chapter, where we will get a grasp of how we can approximate our action-value function and give our agent the power to generalize its experiences.

Chapter 2

Deep Reinforcement Learning

While traditional RL methods, e.g. the Q-learning algorithm, have nice properties with respect to stability and convergence, they're not always applicable when dealing have high-dimensional sensory inputs. Imagine that we want to create an agent with the purpose of playing 64-pixel arcade games, simply by "looking at" and analyzing the screen. This means that the input, i.e. the states, have dimensionality on the order of $64 \times 64 = 4096$. Hence, the number of unique states in the state space, $\mathcal{S}$, is potentially enormous, which means that Q-learning is likely to be computationally infeasible due to the number of state-action pairs.

Since the 90s, attempts have been made to find a more slick solution to estimating the value of action-value pairs in systems with large state-spaces. To date, the most prominent of these attempts is arguably the combination of deep learning and RL through the approximation of the action-value function with a neural network. This idea culminated in the projects by Google DeepMind, presented in Mnih et al. (2012, 2015), in which an RL agent, combined with a deep convolutional neural network, exceeded human-level performance in a number of arcade games by representing the state with 210×160 RGB visual inputs, i.e. dimensionality corresponding to $210 \times 160 \times 3 = 100800$.

2.1 Q-learning Powered by Deep Learning

While the approximation mechanism in a Q-learning algorithm is strictly local, i.e separate estimates for each state-action, the approximation mechanism in a deep neural network is global. In other words, combining an RL agent with a deep neural network gives it the ability to generalize its estimates. To consider an example, let's go back to our favorite biking kid. Imagine that the kid leans too much to the right, causing the whole bike to fall over and hit the ground. Now, if our young biker's brain was wired like a Q-learning algorithm, the fall to the right, and the pain that came with it, would say nothing about what would happen in a similar situation where the bike was instead tilted to the left. This is of course absurd, and we should feel lucky we don't have Q-learning algorithms running the brain department, because the human brain has a powerful ability to generalize and draw comparisons. This is essentially the ability we want to give our RL agent; instead of having to thoroughly experience everything, we want it to be able to use limited experiences to get a comprehensive, general understanding of the environment.

One of the predecessors of DeepMind's arcade-game master was presented by Reidmiller (2005) under the name of *neural-fitted Q-iteration*. Reidmiller presented the problem as finding a tool to balance the positive and negative effects of using a global approximation, rather than a local one. While a global approximation might nullify the training from an older experience when adjusting the approximation based on a recent experience, it can also accelerate training significantly by exploiting generalization. The principal method proposed to achieve this goal is to store experiences and reuse them whenever the approximation of the Q-function is updated. This is based on the idea of *experience replay*, presented by Lin (1992)

In the previous chapter the update rule for the Q-learning algorithm was used, which was based on making incremental updates to the action-value estimates using the Bellman optimality equation:

$$Q_{n+1}(S_t, A_t) = Q_n(S_t, A_t) + \alpha \left[R_{t+1} + \gamma \max_a Q_n(S_{t+1}, a) - Q_n(S_t, A_t) \right]$$

Now, we consider similar update rule, but with a Q-function approximated by a neural network and parametrized by a set of weights, θ, such that we have $Q_n(s, a) = Q(s, a : \theta_n)$ and want to make updates by minimizing a loss:

$$\left(\left[r + \gamma \max_{a'} Q(s', a'; \theta_n) \right] - Q(s, a; \theta_n) \right)^2$$

with respect to the set of weights, θ_n, using e.g. a stochastic gradient descent (SGD) algorithm on previous experiences. In other words, we're concerned with finding a set of weights, θ, such that $Q(s, a; \theta) \approx Q_*(s, a)$. Reidmiller used this technique successfully on a number of simple control problems where the neural-fitted Q-algorithm found good policies relatively fast, compared to analytical model-based techniques. However, it was the introduction of two additional algorithmic features by the DeepMind team that eventually created the Atari-playing RL agent with superhuman game performance.

2.2 The Deep Q-Network

In 2012, a group of researchers from DeepMind, including the aforementioned Martin Reidmiller, released a paper which presented an algorithm that could successfully learn how to play a number of Atari games using a Q-learning agent powered by a deep, convolutional neural network (Mnih et al, 2012). The algorithm, called *Deep Q-Learning*, is very similar to Reidmiller's neural-fitted Q-iteration, except that it uses a convolutional neural network and a more efficient type of experience replay. It also incorporates a so-called *target network*, which is used to train the local network, i.e. the one used for decision making. The weights of the local network are then copied on to the target network at some pre-determined frequency.

Similarly to Reidmiller's algorithm, Deep Q-Learning fits an estimator to the Bellman optimality equation using the loss function

$$L_n(\theta_n) = \mathbb{E}_{s,a\sim\rho}\left[y_n - Q(s,a;\theta_n)\right]^2$$

where $\rho(s,a)$ is a probability distribution over states, s, and actions, a, and

$$y_n = \mathbb{E}_{s'}\left[r + \gamma \max_{a'} Q(s',a';\theta_n^-)\middle| s,a\right]$$

where θ^- are the weights of the target network. Instead of using the entire replay memory, Mnih et al. (2012) take a mini-batch of samples from the experience replay memory and perform SGD, minimizing $L(\theta)$ with respect to θ using these samples. This is done at every step of the algorithm, using the target network. When a deep convolutional neural network is used, the estimator of Q is called a *Deep Q-Network* (DQN).

In 2015, Mnih et al. released another paper where the DQN was applied to 49 different Atari games. The agent achieved more than 75% of human score in more than half of the games, as well as beating humans in several games. To give an idea of how the problem was set up, and how the problem in this thesis will be set up, the authors used, for example, a mini-batch size of 32, a replay memory size of 100000, a target network update frequency of 10000, and a discount factor, γ, of 0.99.

2.2.1 Experience replay

The notion of experience replay is in no way new to Mnih et al. (2012, 2015), but the authors find a more efficient use of it by combining it with random sampling and SGD. However, the use of experience replay is primarily due to stability. In games, and control problems in general, certain states are often interconnected, meaning that they are correlated and hence not independent (Mnih et al, 2012). In other words, a bias will appear when the agent is learning. Sampling from the replay memory remedies this by continuously letting the agent re-experience old state-action pairs. More specifically, we never let the agent improve it's estimations with immediate experiences; we always sample from the replay memory when updating the network, at every step. Reidmiller (2005) noted significant improvements in stability and training time from using the replay memory, making the learning process more stable and data efficient.

2.2.2 Target network

Together with experience replay, the target network is what really makes the deep Q-network an efficient RL agent. When updating the decision-making network at every time-step, the risk of policy divergence and instability increases. If an update increases $Q(s_t, a_t)$, it is likely that $Q(s_{t+1}, a)$ also increases for all a, even though it's not a good estimate of the optimal policy. Similarly to experience replay, the target network makes the learning process more stable and efficient. Mnih et al. (2015) show the effect of using experience replay and a target network for a number of games. The effect on the agent's performance is astounding and while experience replay accounts for the greatest part, the use of a target network improves the performance of the agent many times over in several cases.

2.3 Summary

As mentioned previously, we refrained from discussing convergence and stability properties of Q-learning since the method we'd be using cannot make guarantees on convergence. Boyan and Moore (1995) were early to discuss this problem and showed that the combination of function approximation and certain RL methods could lead to serious instabilities and bad policies. It's evident that this problem is still pervasive, but in using the contributions by Lin (1992), Reidmiller (2005), and Mnih et al. (2015), we arrive at the DQN which does a good job in maintaining both stability and efficiency. With exception for the convolutional neural network used by Mnih et al. (2012, 2015), this is the approach which will be followed in this thesis.

Chapter 3

Real-Time Bidding

During recent years, online display advertisement has been revolutionized by a process now referred to as *Real-Time Bidding* (RTB). Zhang, Yuan and Wang (2014) provide a great introduction to RTB. Rather than advertisers buying specific keywords in search engines or certain fixed time slots on websites, many websites and advertisers have now instead turned to real-time auctions where the display advertisement slot is sold while a website is loading. Due to cookie technology, this means that advertisers can target internet users, or *impressions*, in specific demographic groups and direct their messages to the most appropriate group of people - when they want.

In terms of efficiency, this is quite the innovation. When advertising was dominated by newspaper ads and billboards, an advertiser would basically have to target a very large group, and only under a limited amount of time (as billboards and, especially, newspaper ads could be expensive). With RTB, advertisers can spend their budgets only on the type of people and impressions they actually want to reach.

However, there is a constraint: the process of auctioning out an ad slot and subsequently posting the ad from the successful advertiser often takes less than 100 milliseconds. To put this in perspective, blinking your eye takes about 300 to 400 milliseconds. Hence, buying impressions through RTB auctions is a purely algorithmic procedure and it's easy to see how RL fits into such an algorithmic setting: we have an agent which we want to maximize a reward, i.e. the number of impressions or clicks, under some constraint, i.e. the campaign budget, given a finite set of states and actions. The rest of this chapter will be devoted to describing how a bidding agent can be modeled using RL and, specifically, using the DQN described in the previous chapter.

3.1 Modeling a bidding agent

3.1.1 Considering the framework in chapter 1

For example, how do we handle transition probabilities?? What does the action-value function look like?

3.2 Reinforcement Learning in Real-Time Bidding

3.2.1 Model-based approach

3.2.2 Model-free approach

Chapter 4

Method

4.1 Setting up the problem

Describe what was needed to actually consider the problem and experiment, i.e. in terms of programming. Mention that everything had to be reproduced from scratch. Specify use of tensorflow and so on.

4.1.1 Building the agent

Describe building the agent, in terms of programming, i.e. class-based solutions and how these were integrated into one agent.

4.1.2 Mountain Car test

4.1.3 Building the environment

Describe how environment was built and why.

4.2 Training the agent

Due to the lack of clear examples, I had to improvise here. I partitioned the iPinYou-campgains into smaller episodes and initialized the budgets as random variables with a mean corresponding to the share of the campaign-specific budget for those impressions, with some variance. Normal distribution with unit variance?

MAKE REWARD FUNCTION that incorporates BOTH CTR estimation and actual click??

4.2.1 Would it be better to just shuffle all training data??

4.2.2 Consider somehow improving the use of the replay memory in order to save important experiences, due to massive amounts of data?

4.2.3 Computational challenges, parallel computations, etc

Describe the challenge of not using parallel computing, especially w.r.t. runtime.

4.2.4 Serial training and testing vs. pairwise training and testing

Discuss using training from all campaigns and then testing on all campaigns versus training and testing separately on each campaign. This is a case of expedience and accuracy versus generality.

4.3 Reproducing Wu et al. (2018)

Describe hyperparameters from the paper and how the agent and problem was used together with these to reproduce the model in the paper. Mention not attempting (yet) to reproduce RewardNet and adaptive ϵ-greedy policy. Also mention unclarity around budget etc.

4.4 Data

4.4.1 iPinYou

COULD use processed data from both Cai et al. (2017) AND Du et al. (2017)?? Discuss the data in Wu et al. (2018) as well.

4.4.2 Adform

What's in the Adform data? Is it the same as iPinYou or different in any way, e.g. w.r.t. the CTR estimations???

4.5 Modeling bias and constraints

Check the article sent by Morten!! Perhaps this should be in the discussion/results section??

4.6 Stability testing

Run simulations with exact same hyperparameters and compare results, using some metric, e.g. clicks won or winning rate. Then vary the hyperparameters and see how these changes affect the outcomes.

Chapter 5

Experiments

5.1 Importance of budget

It could be interesting to test the variance of the reward and win rate with different budgets, i.e. to check how an increased budget influences the results, if the relationship is linear, "logarithmic", etc.

5.2 Initializing the λ-parameter

Here, we want to discuss how the performance changes when the Lambda is initialized in certain ways, especially if the Lambda is initialized randomly with some consideration of the historical ratio between winning bids and CTR estimations.

5.3 RewardNet and reward functions

We should consider creating a reward function which incoroprates both CTR-value and eCPC. The reward would be proportional to CTR and inverse proportional to eCPC.

5.4 ϵ-greedy policies

5.5 Dueling Deep Q-Networks

Chapter 6

Results

References

Appendix A

Proofs

Appendix B

Code