



Initializing deep Q-learning under final-only rewards with smoothed Bellman equations

Adrien Chan-Hon-Tong

► To cite this version:

Adrien Chan-Hon-Tong. Initializing deep Q-learning under final-only rewards with smoothed Bellman equations. 2024. hal-04668403

HAL Id: hal-04668403

<https://hal.science/hal-04668403v1>

Preprint submitted on 6 Aug 2024

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



HAL Authorization

Initializing deep Q-learning under final-only rewards with smoothed Bellman equations

Adrien CHAN-HON-TONG

August 6, 2024

Abstract:

Deep Q-learning may fail on trivial problems like maze exploration under final-only reward. Despite reformulation, reward shaping, curiosity exploration or alternative reinforcement algorithms (like proximal policy optimization) may bypass such issue, this paper remarks that finding a path to a final positive state can be done without randomness by smoothing Bellman equations in typical settings. This can be a critical initialization process before regressing the true Q-values.

1 Introduction

Reinforcement Learning (RL) [2] is the ability to learn a global policy on sequential problem from loose supervision only provided by local transition reward. Despite small size RL problems can be solved exactly by tabulating all states, current researches focus on using large deep networks [3] to model the large/infinite set of states and/or the policy. An example is deep Q-learning [4], where residues to so-called Bellman equation are minimized while exploring.

Formally, $\forall (s, a, r, s') \in S \times A \times \mathbb{R} \times S$ a state s , an action a , a reward r and the state s' reached from s by selecting action a , the expected reward Q^* (discounted by $\gamma \in]0, 1[$) that can be obtained from s, a is

$$Q^*(s, a) = r + \gamma \max_{a'} Q^*(s', a') \quad (1)$$

Thus, if Q is a deep network with weights θ modeling Q^* , then, one can use Q and randomness or diversity to explore, store experiments in replay buffer $\mathcal{R}$ while updating the weights by minimizing the loss

$$l(\theta) = \sum_{(s, a, r, s') \in \mathcal{R}} (r + \gamma \max_{a'} Q(s', a') - Q(s, a))^2 \quad (2)$$

However, a large numbers of papers (e.g. [6]) have proven that it may fail:

- Deceptive rewards may place the convergence on the hand of the greedy epsilon strategy (despite Soft-DQN [5] or exploration based on curiosity [9] or entropy [8] may mitigate this issue).
- As the modeled policy controls the exploration, errors can be amplified (despite double DQN and replay buffer may mitigate this issue [7]).
- Training deep networks can be hard even on well posed problem. For example, [1] proves that deep networks may diverge on some RL setting.

While tackling those problems in general may be a long quest, this paper focuses on a specific RL setting where the following condition holds.

Definition of the considered condition: *all starting states have a positive average Q values and all rewards except when reaching good exit states are negative.*

Obviously, this condition is not intended for general RL but it has several interest. In addition, this condition is somehow natural for example for video game as *people usually do not play game which can not be won*. Precisely, under this strong assumption, this paper proves that preventing locked exploration can be done by smoothing Bellman equations:

- Next section will precise why final-only reward can be deceptive.
- Then, section 3 will prove that adding the constraint $\sum_a Q(s, a) > 0$ on starting states s ensures that optimized Q model is never locked.
- Finally, optimization considerations are tackled in section 4 before conclusion of section 5.

2 Hardness of final-only reward instances

2.1 Deceptive maze exploration

Let consider a very simple RL instance: $A = \{-1, 1\}$, $S = \{-1, 0, 1, \dots, N\}$ with -1 and N being final state, $s' = s + a$ and $r(i, i \pm 1) = -\alpha$ except for $r(N - 1, N) = \mu$ ($\mu \gg \alpha > 0$).

This problem is deceptive because selecting -1 from 0 lead to a total reward of $-\alpha$, but, one needs N more actions 1 than -1 to reach N eventually getting a positive reward while selecting less than N action 1 leads to a smaller reward than $-\alpha$ (up to $-\sum_{i \leq N} \gamma^i \alpha < -\alpha$). Thus, naive exploration will tend to select action -1 from 0 . For example, $Q(-1) = 0$, $Q(0, -1) = -\alpha$ and $Q(0, -1) < -\alpha$ is related to a 0-loss of eq2 if $\mathcal{R}$ contains only 0 and -1 which is the case for a deterministic exploration. Thus, exploring state $n \in \{1, \dots, N\}$ from Q requires random-based or curiosity-based exploration. However, if N is large both randomness or curiosity may fail: reaching N with ε -greedy strategy has probability ε^N .

However, the true Q value of $n \in \{1, \dots, N-1\}$ is given by $Q^*(n, 1) = \gamma Q^*(n+1, 1) - \alpha$ and $Q^*(N-1, 1) = \mu$ i.e. $Q^*(n, 1) = \gamma^{N-n} \mu - (\sum_{i \leq N-n} \gamma^i) \alpha$.

Thus, for sufficiently large μ , $\sum_a Q^*(0, a) > 0$.

2.2 A more general setting

More generally, let S a set of states, A a set of actions, r a reward function (i.e. a function from $S \times A$ to $\mathbb{R}$), one can consider 3 subsets of S , $\mathcal{S} \subset S$ a set of starting states, $\mathcal{F} \subset S$ a set of final states and $\mathcal{G} \subset \mathcal{F}$ a set of good exit states. Let write $T(s, a) \in S$ for the state reached by applying action $a \in A$ from state $s \in S - \mathcal{F}$ (more usually written s'). Let also write $\mathcal{T}(s) \subset S$ the set of states that can be reached by a path from $s \in S$ (i.e. $\mathcal{T}(s)$ is the smallest subset of S such that $\forall \sigma \in \mathcal{T}(s), \forall a \in A, T(\sigma, a) \in \mathcal{T}(s)$). Finally, let write $\pi(s) = \arg \max_a Q(s, a)$.

Formalization of the considered condition: The considered condition is that one has the prior that

1. $\forall s \in \mathcal{S}, \sum_a Q^*(s, a) > 0$
2. $\forall s \in \mathcal{S}, \mathcal{T}(s) \cap \mathcal{G} \neq \emptyset$
3. $\forall s, a \in S \times A, T(s, a) \notin \mathcal{G} \Rightarrow r(s, a) < 0$

Remark 1: Obviously, a more general setting is composition of 2 such instances i.e. states can be split in 2 disjoint sets with good exit states from first being the starting states of the second.

3 Lock-free exploration

Previous section defines properly the condition considered in this paper. Then, this section will now investigate the advantage that can be taken from this condition.

3.1 Definitions

0-loss: $\forall Q$ a function of $S \times A$ to $\mathbb{R}$, the function is called 0-loss on $\Gamma \subset S$ if it verifies the Bellman equation on Γ (more precisely a slightly weaker version where the equation should only hold for maximal action):

$$\begin{aligned} \forall s \in \Gamma \cap \mathcal{F}, \forall a \in A & \quad Q(s, a) = 0 \\ \forall s \in \Gamma - \mathcal{F}, & \quad Q(s, \pi(s)) = r(s, \pi(s)) + \gamma Q(T(s, \pi(s)), \pi(T(s, \pi(s)))) \end{aligned} \quad (3)$$

Locked exploration: Let now define a locked exploration as a situation where there exists $\Gamma \subset \mathcal{S}$ and Q a 0-loss function on Γ with the additional properties:

1. $\mathcal{S} \subset \Gamma$
2. $\Gamma \cap \mathcal{G} = \emptyset$
3. $\forall s \in \Gamma, T(s, \pi(s)) \in \Gamma$

Importantly, under the assumptions that there is a locked exploration, then, all the states that will be reached by a Q -path starting from $\mathcal{S}$ will be contained in Γ and thus never reach $\mathcal{G}$ (as $\Gamma \cap \mathcal{G} = \emptyset$). This statement can be easily proven by induction: this is true for initial point as $\mathcal{S} \subset \Gamma$, then, considering $s \in \Gamma - \mathcal{F}$, the next state induced by Q is $T(s, \pi(s))$ which is in Γ according to the assumptions 3 of the locked exploration.

Thus, the phrase *locked assumptions* is consistent with general understanding. Despite ε -greedy exploration or curiosity may theoretically mitigate this issue, the maze example of 2.1 recalls that it is not acceptable to have only ε^N probability to exit. This invites to consider other solutions to locked exploration.

Main theorem:

If $\forall s \in \mathcal{S}, \max_{a \in A} Q(s, a) > 0$, then, no situation of locked exploration can happen.

Proof: let proceed by assuming there is a locked exploration and proving a contradiction. Let first consider $\Omega = \max_{s \in \Gamma, a \in A} Q(s, a) > 0$. As, from assumption 1 of locked exploration, $\mathcal{S} \subset \Gamma$, it means that $\Omega \geq \max_{s \in \mathcal{S}, a \in A} Q(s, a)$.

But, the considered condition is that $\max_{s \in \mathcal{S} \cap \mathcal{G}, a \in A} Q(s, a) > 0$, so $\Omega > 0$.

Let now consider s, a such that $Q(s, a) = \Omega$. By definition of the max, $a = \pi(s)$. From assumption that Q has 0-loss on Γ (eq3), $Q(s, a) = r(s, a) + \gamma Q(T(s, a), \pi(T(s, a)))$. But, from the assumption 3 that the exploration is locked, $T(s, a) \in \Gamma$ so $Q(T(s, a), \pi(T(s, a))) \leq \Omega$.

So, $\Omega < r(s, a) + \gamma\Omega$. But, by assumption 2 of the locked exploration, $T(s, a) \notin \mathcal{G}$ which means that $r(s, a) < 0$ seeing condition 3 on the instance.

So, $\Omega < \Omega$. This is a contradiction.

Remark 2: So, the assumption $Q(s, a) > 0$ for $s \in \mathcal{S}, a \in A$ and 0-loss on eq3 imply that the exploration is never locked: minimizing the equation 3 with positivity constraint will eventually lead to a policy such that $s = T(s, \pi(s))$ terminate with $s \in \mathcal{G}$. As, the existence of a 0-loss function Q on $\mathcal{S}$ is a well known result related to Bellman equation, solving equations 3 with constraint $Q(s, a) > 0$ is both possible on the considered RL instances (let recall that assumption 2 states $\mathcal{T}(s) \cap \mathcal{G} \neq \emptyset$) and provides a lock-free solution.

4 Practical optimization

Despite previous section has proven that the addition of the simple constraint $\forall s \in \mathcal{S}, \sum_a Q(s, a) > 0$ guarantees that any 0-loss model is lock-free on instance verifying the considered condition, it worths remarking that from gradient based optimization point of view the convergence of a deep network toward a loss level of 0 is not trivial.

Indeed, in the maze example, the situation $Q(0, 1) = 1, Q(1, -1) = 1$ is a local minima of the system $Q(0, 1) \geq 1, Q(0, 1) = -\alpha + \gamma Q(1, -1)$ and $Q(1, -1) = -\alpha + \gamma Q(0, 1)$. Indeed, the gradient when to decrease equally $Q(0, 1)$ and $Q(1, -1)$, but the constraint $Q(0, 1) \geq 1$ forbids it. Thus, so optimization is locked resulting in a suboptimal model unable to find a good exit (and related to a high Bellman loss).

Yet, by replacing $\max_a Q(s, a)$ by $\frac{\Upsilon \max_a Q(s, a) + \sum_a Q(s, a)}{\Upsilon + |A|}$ with positive Υ , one can smooth the optimization while maintaining a the property of avoiding a locked exploration.

Indeed, let assume that $\mathcal{Q}$ is positive on $\mathcal{S}$ states and verifies on Γ :

$$\mathcal{Q}(s, a) = r(s, a) + \gamma \frac{\Upsilon \max_{a'} \mathcal{Q}(T(s, a), a') + \sum_{a'} \mathcal{Q}(T(s, a), a')}{\Upsilon + |A|} \quad (4)$$

then, the same conclusion than in the main theorem holds: $\Omega = \max_{s \in \Gamma, a \in A} \mathcal{Q}(s, a) > 0$ and $\Omega = r(\sigma, \alpha) + \gamma \Omega$ where σ, α is the argmax related to Ω , so $r(\sigma, \alpha) > 0$, so $T(\sigma, \alpha) \in \mathcal{G}$.

Thus, any models with a 0-loss for the smooth Bellman equations (eq4) can not be locked from theoretical point of view and the optimization should be much smoother thank to the $\sum_{a \in A} \mathcal{Q}(s, a)$ in eq4 which makes gradient flow to reach each pair s, a .

The limit of this approach is that a model $\mathcal{Q}$ with 0-loss for eq4 can not be interpreted as a Q value i.e. it does not approximate the expected discounted reward that can be obtained from s, a , and, worse, it may have an effect similar to a modification of γ to $\frac{\Upsilon+1}{\Upsilon+|A|}\gamma$: if for a state $T(s, a) = s', \mathcal{Q}(s', a')$ is constrained to be 0, except for $a' = \alpha$, then, $\mathcal{Q}(s, a) = r(s, a) + \gamma \frac{\Upsilon+1}{\Upsilon+|A|} \mathcal{Q}(T(s, a), \alpha)$ exactly if γ was $\gamma \frac{\Upsilon+1}{\Upsilon+|A|}$. However, it may be relevant to initialize a deep Q network.

Remark 3: The case $\Upsilon = 0$ allows to prove easily the existence of a 0-loss model. Indeed, if one can tabulate all states, then (when $\Upsilon = 0$), eq4 is a linear system of all states. Let assume that lines are organized so that if i represents s, a , then, constraint $\mathcal{Q}(s, a) = r(s, a) + \gamma \sum_{a'} \mathcal{Q}(T(s, a), a')$ is in line i . Then, one can write the matrix of this system $I - \Theta$ where Θ has the property that $\max_i \sum_j |\Theta_{i,j}| = \gamma < 1$. But $\max_i \sum_j |\Theta_{i,j}|$ is larger than the larger eigen value. Indeed, $\lambda v = \Theta v \Rightarrow \lambda v_i = \sum_j \Theta_{i,j} v_j$. So $\lambda \max_k v_k \leq \max_k v_k \max_i \sum_j |\Theta_{i,j}| \leq \max_k v_k$ i.e. $\lambda \leq \max_i \sum_j |\Theta_{i,j}| < 1$. Thus, $I - \Theta$ is not singular i.e. there exists a model $\mathcal{Q}$ with 0-loss.

However, it is related to a $\frac{\gamma}{|A|}$ decay which is not convenient. Inversely, $\Upsilon > |A| - 1$ guarantees that the sign of the max is the sign of its smoothed version, and, proof of existence is also related to the fact that $\gamma < 1$ implies that the operator which moves a distribution on S to the distribution after transition is a contraction mapping. Anyway, constraining Q to be lower bounded can be relevant to avoid divergence on unconstrained pair s, a during optimization in practice.

5 Conclusion

This paper considers the RL instances where positive rewards can be obtained only when reaching some final states (or composition of such instances) e.g. maze exploration with large positive reward on exit.

Despite this setting is obviously restricted, this paper remarks that training a deep network related to a policy which finds an exit can then be done without randomness nor curiosity just by smoothing Bellman equation and adding positivity constraint on initial states: indeed, considering the positive maximal Q value of set of states in this setting trivially shows (under 0-loss assumption) that either the next one is out of the set or is a positive final one, so exploration can not be locked. From such initialization, one can refine a classical deep Q-learning with a replay buffer containing path to a good exit.

Future works should consider practical experimentations of this framework.

References

- [1] Joshua Achiam, Ethan Knight, and Pieter Abbeel. Towards characterizing divergence in deep q-learning. *arXiv preprint arXiv:1903.08894*, 2019.
- [2] Leslie Pack Kaelbling, Michael L Littman, and Andrew W Moore. Reinforcement learning: A survey. *Journal of artificial intelligence research*, 4:237–285, 1996.
- [3] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 2015.
- [4] Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Alex Graves, Ioannis Antonoglou, Daan Wierstra, and Martin Riedmiller. Playing atari with deep reinforcement learning. *arXiv preprint arXiv:1312.5602*, 2013.
- [5] Zhao Song, Ron Parr, and Lawrence Carin. Revisiting the softmax bellman operator: New benefits and new perspective. In *International conference on machine learning*, pages 5916–5925. PMLR, 2019.
- [6] Hado Van Hasselt, Yotam Doron, Florian Strub, Matteo Hessel, Nicolas Sonnerat, and Joseph Modayil. Deep reinforcement learning and the deadly triad. *arXiv preprint arXiv:1812.02648*, 2018.

- [7] Hado Van Hasselt, Arthur Guez, and David Silver. Deep reinforcement learning with double q-learning. In *Proceedings of the AAAI conference on artificial intelligence*, volume 30, 2016.
- [8] Markus Wulfmeier, Peter Ondruska, and Ingmar Posner. Maximum entropy deep inverse reinforcement learning. *arXiv preprint arXiv:1507.04888*, 2015.
- [9] Oleksii Zhelo, Jingwei Zhang, Lei Tai, Ming Liu, and Wolfram Burgard. Curiosity-driven exploration for mapless navigation with deep reinforcement learning. *arXiv preprint arXiv:1804.00456*, 2018.