We thank the reviewers for all constructive reviews and will correct all minor problems accordingly .

Reviewer 1: Thanks for the comments. **(1)** We add a new experiment on the traffic light control problem [4] according to requirement of the reviewer. We compare value propagation with MA-AC, independent Q learning, PCL without communication and value propagation with partial observation. See more details in the left panel of Figure 1. **(2)** We introduced our setting (goal of our algorithm) in the introduction session (lines 39 to 43). We will highlight that in the final version so that readers can understand it clearly. **(3)** We have tested our claim on data efficiency of the off-policy method in the experiment. One baseline MA-AC is an on-policy algorithm, and our value propagation algorithm outperforms that due to better data efficiency. **(4)** Thanks for the suggestion on the structure of the paper, we will rectify it in the final version. **(5)** We summarized our main contribution in the contribution section (row 65-76).

Reviewer 2: Thanks for the comments. One possible real-application is the multi-agent autonomous driving system [1], where each agent may not want to share its goal (reward), its driving policy and its evaluation on current condition (value function) . However they still can cooperate to guarantee the safety of driving, e.g., giving way, through the local communication induced by the communication graph.

Reviewer 3: Thanks for the comments. **(1)**. We run the cooperative navigation task with more episodes in the middle (8 agents) and right panel (16 agents) of Figure 1 with an additional baseline independent Q learning. We add a new experiment on the traffic light control in the left panel. **(2)**. The high-level explanation on the proof of convergence. Most commonly used TD algorithms are **semi-gradient** algorithms [2], e.g., Q learning, SARSA, Q and V update in actor-critic. When they optimize the squared TD-error, they do not calculate gradient w.r.t. the parameter $\theta$ of the target (e.g. target in Q learning is $R + \gamma \max_a Q_\theta(s_{t+1}, a)$.). That is why they are so-called semi-gradient algorithm. It would have the convergence problem combining with the function approximation, and off-policy learning (called deadly triad in [2]). One way to survive is to use the **true gradient** method [2], such as Gradient TD (but it is just a policy evaluation method). Value propagation is a true gradient method. It optimizes the objective function in row 100 (single agent case) and eq (8) in multi-agent setting. In optimization theory, even the objective function is non-convex, we still can design some gradient-based algorithms which converge to the stationary point. **(3)** The reason to use primal-dual form. If we directly optimize the primal error (e.g., the objective function in row 100), it would meet the double sampling problem [3]. To get around this problem, value propagation introduces a dual variable. The high-level idea to use the dual variable is similar to Gradient TD (see the w variable in [3]), but now we solve a much harder control problem ( Gradient TD is designed for policy evaluation ). **(4)** Intuition on why value propagation is better than MA-AC. MA-AC is on-policy, which requires new samples to be collected for each gradient step. This becomes expensive, as the number of gradient steps and samples per step needed to learn an effective policy increases with task complexity. Value propagation (off-policy) reuses past experience. In algorithm 1, to update dual or primal problem, we randomly sample mini-batch of transition from the replay buffer. The reason that value propagation is off-policy is proposition 1, which says for all pair (s,a), the temporal consistency holds.

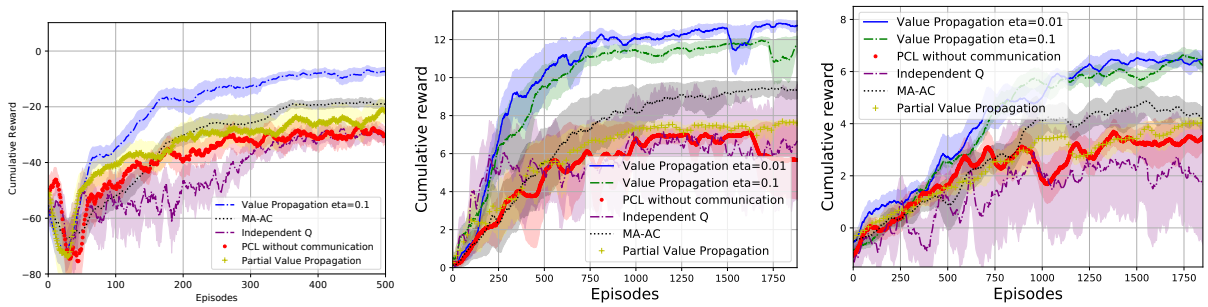


Figure 1: Additional experiments. Left: traffic light control. There are 9 agents where each of them represents a traffic light. Actions are phase transition of traffic light. Rewards are combinations of delays, waiting time, emergency stops of vehicle. Middle and Right panel: cooperative navigation task with more episodes. We add a new baseline, independent Q learning, in the experiment. Value propagation clearly outperforms MA-AC, PCL without communication, and independent Q learning.


[1] Safe, Multi-Agent, Reinforcement Learning for Autonomous Driving. S Shalev-Shwartz et al. 2016

[2] Introduction to Reinforcement Learning with Function Approximation, Richard Sutton, Nips 2015 tutorial.

[3] Fast Gradient-Descent Methods for Temporal-Difference Learning with Linear Function Approximation. Richard Sutton et al. ICML 2009

[4] Coordinated Deep Reinforcement Learners for Traffic Light Control. Elise van der Pol et al., Nips 2016 workshop