A EFFECT OF REWARD SHAPING

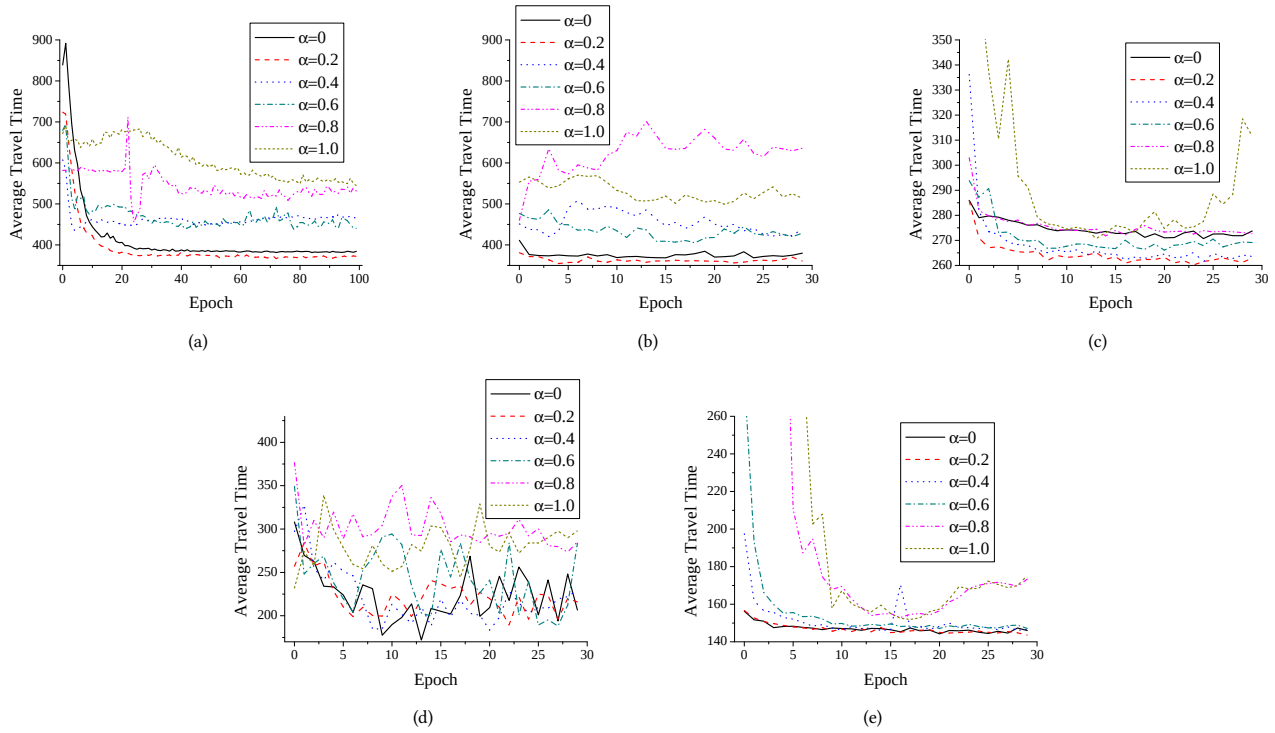

Figure A1: Convergence curves by setting different weight α in the meta-training step and the fine-tuning stage. (a) The meta-training step on $Hangzhou_{4\times4,dis=0}$. (b) The fine-tuning stage on $Hangzhou_{4\times4,dis=0.1}$. (c) The fine-tuning stage on $Jinan_{3\times4,dis=0}$. (d) The fine-tuning stage on $Atlanta_{1\times5,dis=0}$. (e) The fine-tuning stage on $Syn_{3\times3,dis=0}$.

In Meta-CSLight, the reward shaping can enhance the traffic efficiency by encouraging more green waves. This paper uses the weight $\alpha \in [0, 1]$ to balance the immediate reward and the reward shaping. To investigate the sensitivity of α, in the meta-training step and the fine-tuning stage, the convergence curves are plotted by setting $\alpha = 0, 0.2, 0.4, 0.6, 0.8, 1$, respectively. The convergence curves are shown in Figure A1.

The experimental results demonstrate that Meta-CSLight can converge to the best performance by setting $\alpha = 0.2$. In the meta-training step, in terms of average travel time, Meta-CSLight can converge to the best results in 20 episodes on $Hangzhou_{4\times4,dis=0}$ by setting $\alpha = 0.2$. In the fine-tuning stage, Meta-CSLight can converge to the best results in 5, 23 and 30 episodes on Hangzhou, Jinan and Synthetic datasets by setting $\alpha = 0.2$, respectively. On Atlanta dataset, Meta-CSLight converges to the best result in 13 episodes by setting $\alpha = 0$. However, the convergence curve in last serveral episodes fluctuates wildly. Compared with this curve, Meta-CSLight can stably converge to the similar results by setting $\alpha = 0.2$. It indicates that the reward shaping can improve the convergence of Meta-CSLight on these datasets.

Besides, when $\alpha \geqslant 0.4$, it is hard to converge to better results for Meta-CSLight on these datasets. Maybe the traffic signals of the neighborhood of each agent will make the TSC decision unbalanced by selecting a larger α.

B CASE STUDY

To further show the effect of the reward shaping, this section compares the number of green waves by using the reward shaping with not using it on $Hangzhou_{4\times4,dis=0}$ in the meta-training step. The experimental results are shown in Figure B2.

It can be observed that Meta-CSLight with reward shaping achieves the highest number of green waves in 20 episodes. There are no noticeable increments after 20 episodes. When the reward shaping is removed from Meta-CSLight, in terms of the number of green waves, the performance is similar with the former in the beginning but slowly increases after 20 episodes. It indicates that the reward shaping can accelerate the convergence speed in the meta-training step.

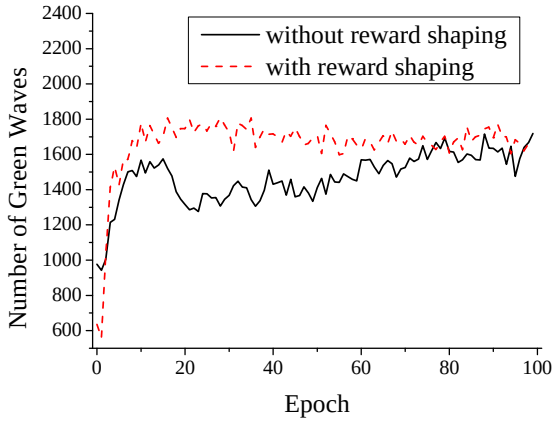

Figure B2: The number of green waves with and without reward shaping on $Hangzhou_{4\times4,dis=0}$ in the meta-training step.