ДЗ5

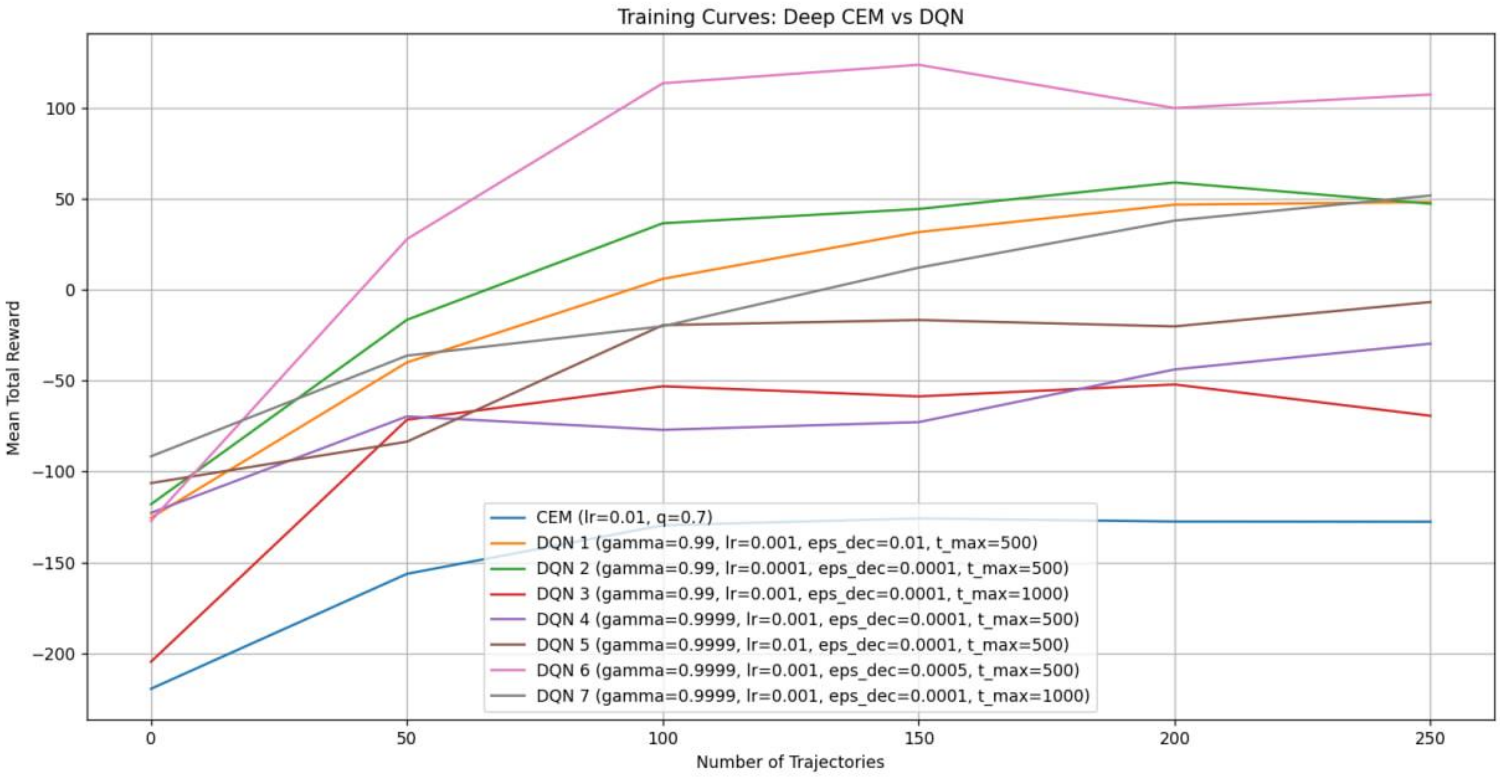
I. Реализация DQN для LunarLanderV2. Сравнение с Deep CEM.

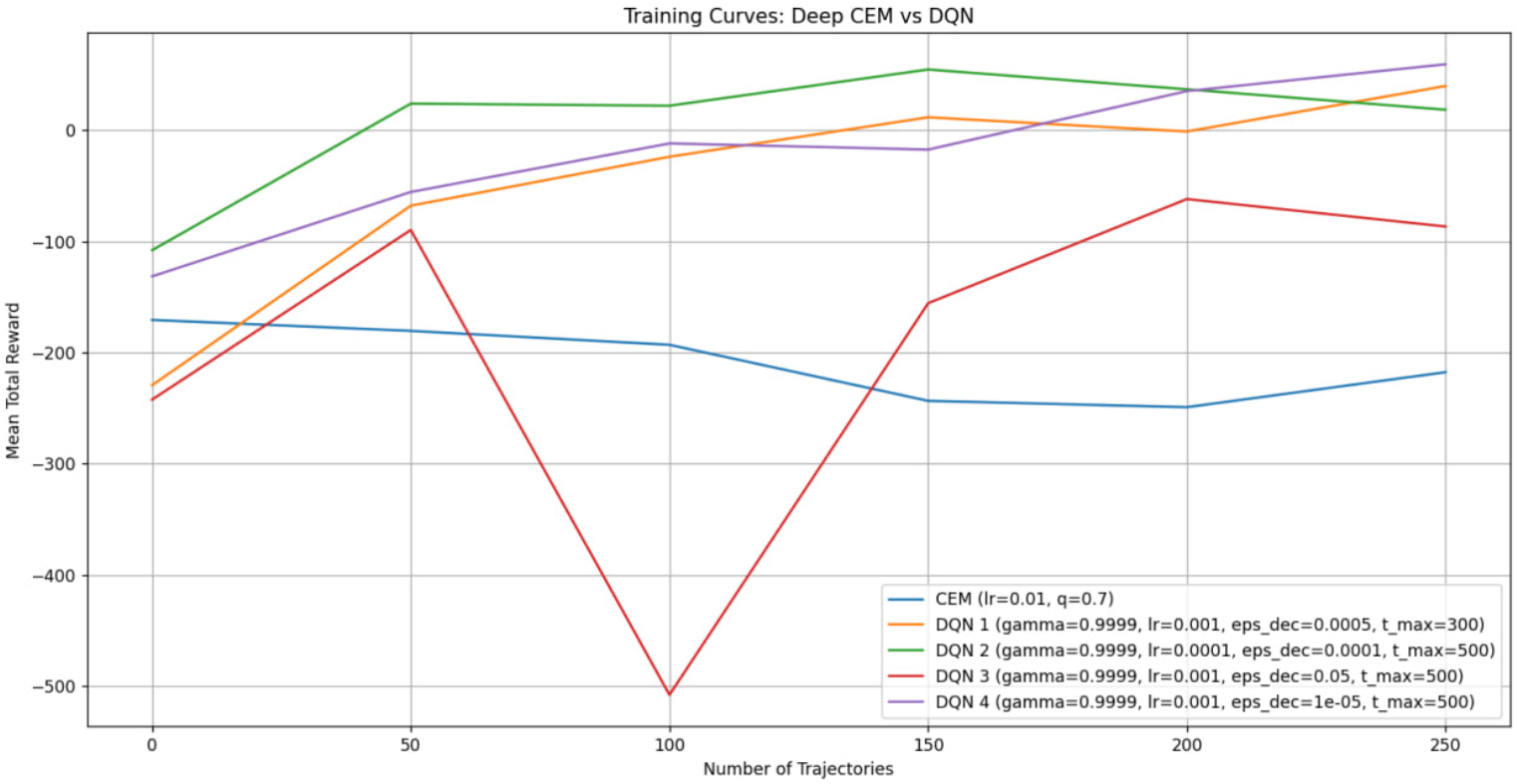
гиперпараметры:

- **gamma**: коэффициент дисконтирования;
- **learning_rate**: скорость обучения нейронной сети;
- **epsilon_decrease**: скорость уменьшения ε в стратегии ε-жадности;
- **t_max**: максимальная длина эпизода.

Метрика оценки: средняя награда за последние 50 траекторий.

Сравнение: оценка производительности алгоритмов по улучшению средней награды и стабильности результатов.





Выводы:

1. **Deep CEM** показывает стабильные, но ограниченные улучшения и достигает плато награды около -127 к 200 траекториям.
2. **DQN** демонстрирует явное превосходство, достигая награды более 100 при оптимальных гиперпараметрах.
3. Наиболее эффективная конфигурация для DQN:
   {"gamma": 0.9999, "learning_rate": 1e-3, "epsilon_decrease": 5e-4, "t_max": 500}

II. Модификации DQN на LunarLander-v2.

Эксперимент направлен на изучение влияния модификаций DQN (**Hard Target Update, Soft Target Update, Double DQN**) на процесс обучения. Рассматриваем следующие новые гиперпараметры:

- tau (коэффициент для Soft Target Update)

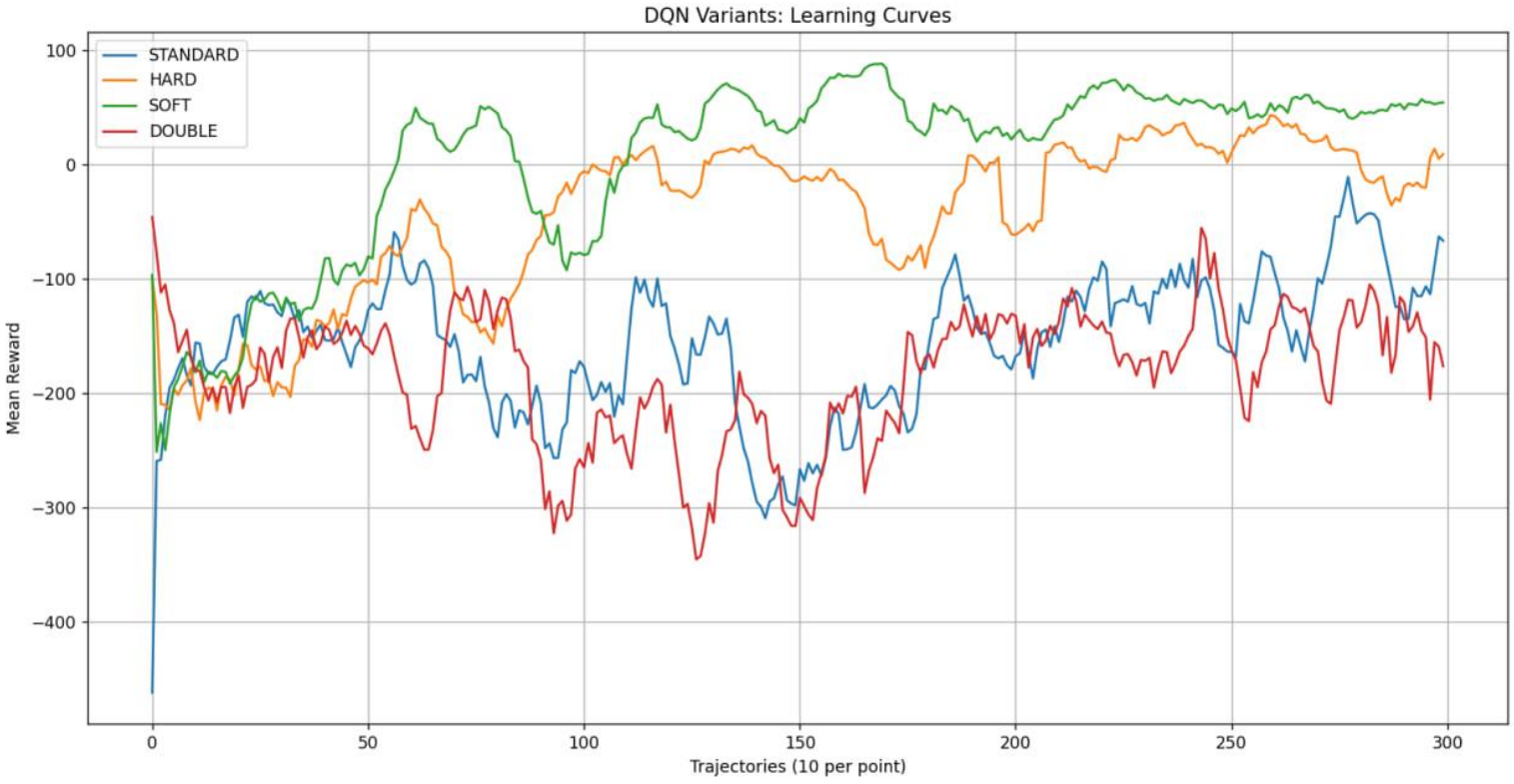
В Soft Target Update используется постепенное обновление параметров целевой сети. Параметры основной сети (Q_main) частично копируются в целевую сеть (Q_target) с учетом коэффициента tau.
Формула обновления: $Q_{target} \leftarrow (1-\tau)Q_{target}+\tau Q_{main}$
При tau = 1 целевая сеть полностью заменяет свои параметры на параметры основной сети, что аналогично Hard Update. При tau = 0 целевая сеть не обновляется.
Для задач, где Q-оценки имеют высокую дисперсию (например, из-за больших вознаграждений), лучше использовать меньшее tau.
Для задач с более предсказуемыми вознаграждениями допустимо большее tau, чтобы ускорить обучение.

-Update Frequency (для Hard Target Update)

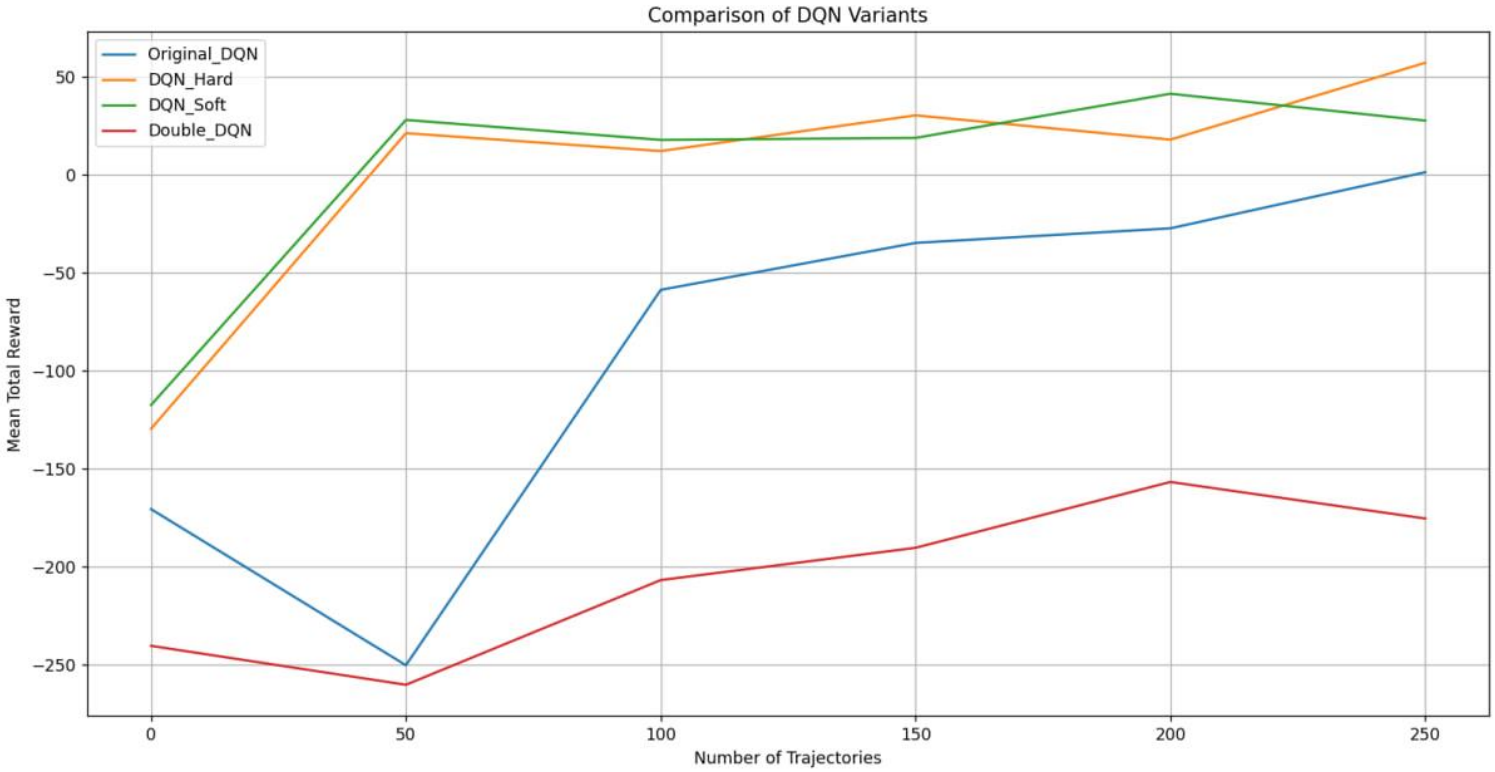
Этот параметр определяет, как часто параметры основной сети полностью копируются в целевую сеть.

Для нестабильных сред или задач с большим пространством состояний лучше использовать более редкое обновление.

Для простых или быстро адаптирующихся задач можно использовать частое обновление, чтобы ускорить обучение.



```
"epsilon_decrease": 1e-4,
"gamma": 0.99,
"batch_size": 64,
"lr": 1e-3,
"epsilon_min": 0.01,
"tau": 0.005,
"target_update_freq": 100,
"total_trajectories": 300,
"t_max": 500,
```



“tau”: 0.7,

“update_freq”: 20

Выводы:

Оригинальный DQN показал прогресс в обучении, однако был менее стабильным и достиг низких средних наград.

Модификации DQN с Hard Target Update и Soft Target Update улучшили стабильность и скорость обучения.
Double DQN не продемонстрировал ожидаемых улучшений.