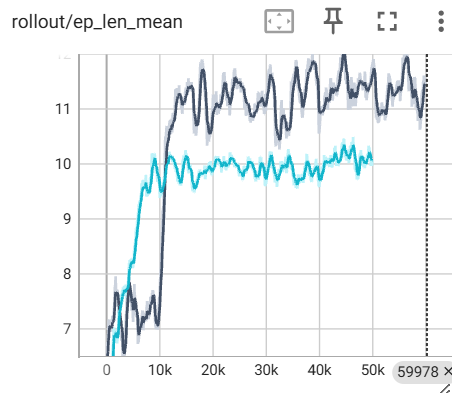
# Merge

A white lines on a gray background

Description automatically generated

The merge-v0 environment is part of the highway-env library, designed to simulate a highway merging scenario. In this environment, the ego vehicle starts on a main highway while other vehicles are on dedicated merging lane and begins to integrate into the main traffic flow on a multi-lane highway. The task requires balancing efficiency and safety, as the agent must avoid collisions and make room for the vehicles while maintaining a reasonable speed and staying within the road boundaries.

**Episode Length Analysis (rollout/ep\_len\_mean)**





The mean episode length is a critical indicator of the agent's ability to navigate the environment without terminating due to collisions or other constraints. From the graph, the **TD3** demonstrates a faster and more consistent improvement in episode length compared to the **SAC model**. TD3 reaches an episode length of approximately **11** by 10,000 timesteps, while SAC lags slightly, reaching similar values later in the training process. Furthermore, the SAC model exhibits more gradual and steady growth, indicating that it learns more conservatively. In contrast, TD3 demonstrates higher variability, which suggests its exploration mechanism may cause greater fluctuations during training.

**Episode Reward Analysis (rollout/ep\_rew\_mean)**

A graph of a graph

Description automatically generated with medium confidence



The mean episode reward reflects the agent's overall performance and ability to optimize the reward function. The **TD3 model** shows a rapid increase in rewards during the initial training phase, achieving a mean reward of approximately 10 by 10,000 timesteps. In contrast, the **SAC model** follows a more gradual trajectory, stabilizing around a mean reward of approximately 9. Unlike TD3, which reaches and maintains higher reward levels, SAC exhibits smaller fluctuations and a smoother learning process. However, SAC consistently falls short of achieving the peak performance demonstrated by TD3, highlighting a gap in its ability to optimize reward as effectively in this environment.

**Actor loss**

A graph with blue lines

Description automatically generated

The **actor loss** provides insights into the optimization process of the policy network in each model. For the **TD3 model**, the actor loss starts at relatively high values and undergoes a significant decrease over the initial 20,000 timesteps, stabilizing around approximately **-5.5**. However, a slight upward trend is observed after 20,000 timesteps, with the actor loss increasing marginally to approximately **-5.2**.

In contrast, the **SAC model** exhibits a similar pattern, with the actor loss decreasing significantly over the first 10,000 timesteps, followed by stabilization. However, SAC shows a slightly higher range of values, decreasing to approximately **-5.0** before increasing slightly to **-4.5**.

These trends indicate that both models initially optimize their policies effectively but show some minor instability in the later stages of training. The slightly higher actor loss in SAC suggests that it maintains a less aggressive policy update compared to TD3, which could contribute to SAC's more stable but less performant behaviour as observed in the reward trends.

**Critic loss**

****

The **critic loss** reflects the efficiency of the value function's optimization, providing insights into how well each model estimates future rewards. For the **SAC model**, the critic loss starts at a relatively high value of over 1.0 and gradually decreases to approximately 0.1 over the course of 50,000 timesteps. This consistent downward trend indicates stable and effective learning of the value function.

In contrast, the **TD3 model** displays a different pattern. The critic loss is not reported during the initial 10,000 timesteps, as TD3 has a delayed training start (learning\_starts=10000) to collect sufficient exploration data before updating the value function. Once training begins, the critic loss is shown to start at approximately 0.12. Over the next 5,000 timesteps, the critic loss rises sharply to over 1.0 before starting to decrease again, eventually stabilizing around 1.0. This sharp increase followed by stabilization suggests that TD3's value function undergoes a period of adjustment after training begins.

These trends highlight the **more stable and gradual learning process of SAC** compared to TD3. SAC's smoother critic loss trajectory complements its observed stability in reward optimization, while TD3's fluctuating critic loss aligns with its more oscillatory reward behavior during training. This suggests that SAC's critic effectively minimizes errors more consistently, whereas TD3 experiences a period of variability before settling.

SAC performance

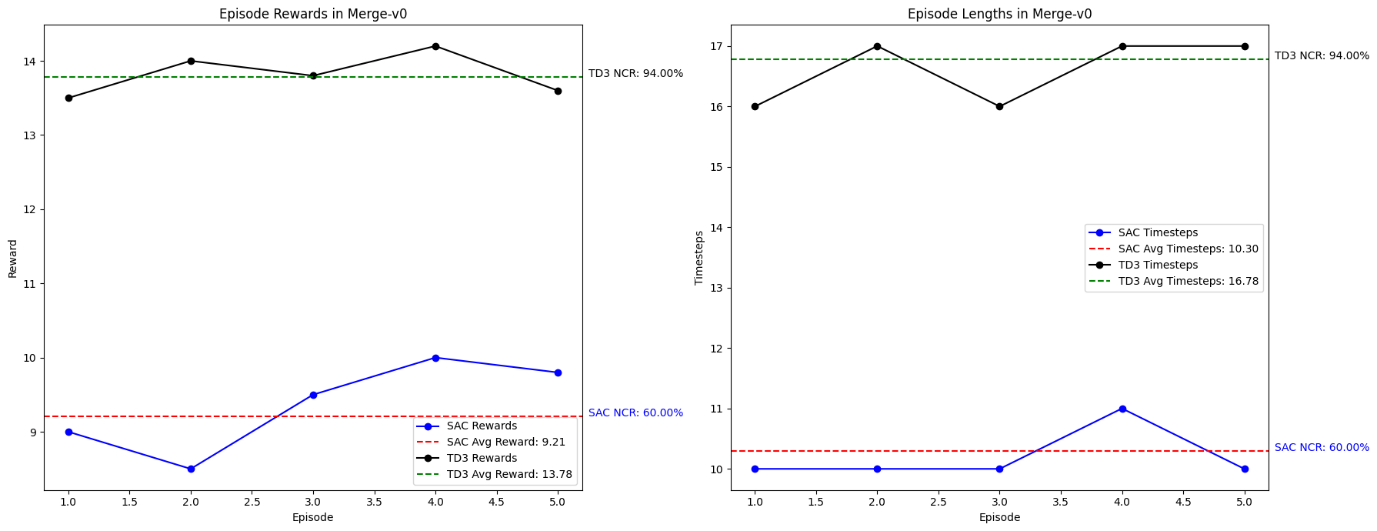
A close-up of a graph

Description automatically generated

A close-up of a graph

Description automatically generated

Test over 100 episodes:



The performance comparison between TD3 and SAC in the merge-v0 environment reveals notable differences in their ability to handle the merging task. TD3 consistently achieves higher rewards across all episodes, with an average reward of 13.78 compared to SAC's 9.21. This performance is further reflected in the Not Crash Rate (NCR), where TD3 achieves 94%, indicating superior stability and reliability in avoiding collisions. In contrast, SAC demonstrates a lower NCR of 60%, suggesting challenges in maintaining safe and efficient control.

Additionally, the episode lengths highlight the models' distinct performance characteristics. TD3 maintains longer average episode durations of 16.78 timesteps, showcasing its ability to remain active in the environment for extended periods and complete the task effectively. Conversely, SAC's shorter episode lengths, averaging 10.30 timesteps, suggest frequent failures or suboptimal decisions, contributing to its lower rewards and success rate. These findings underscore TD3's robustness in managing complex environments like merging scenarios, while SAC exhibits limitations in achieving comparable performance and stability.