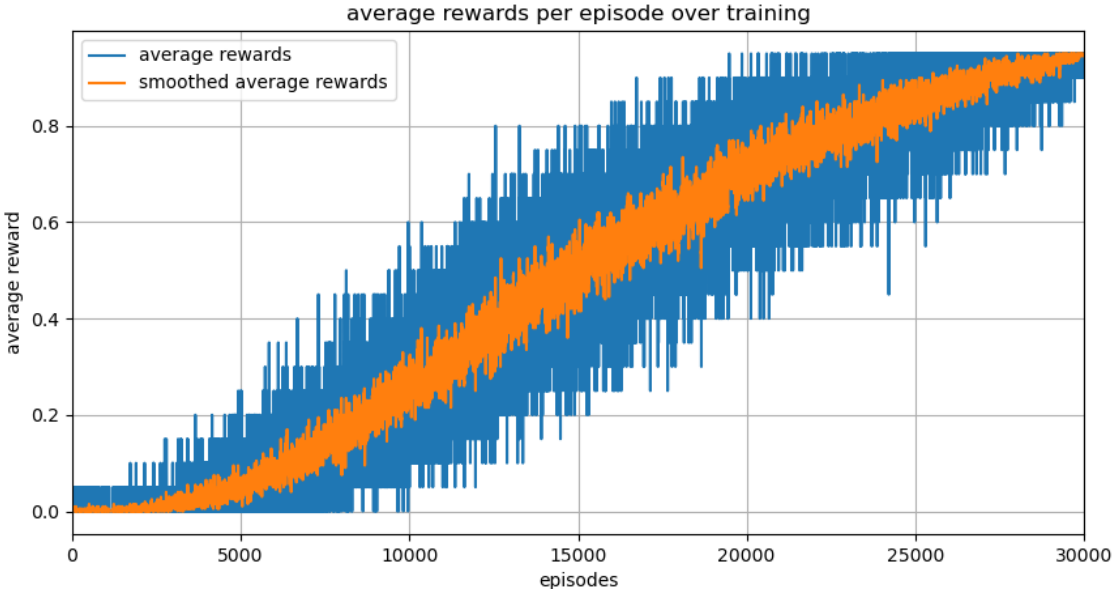
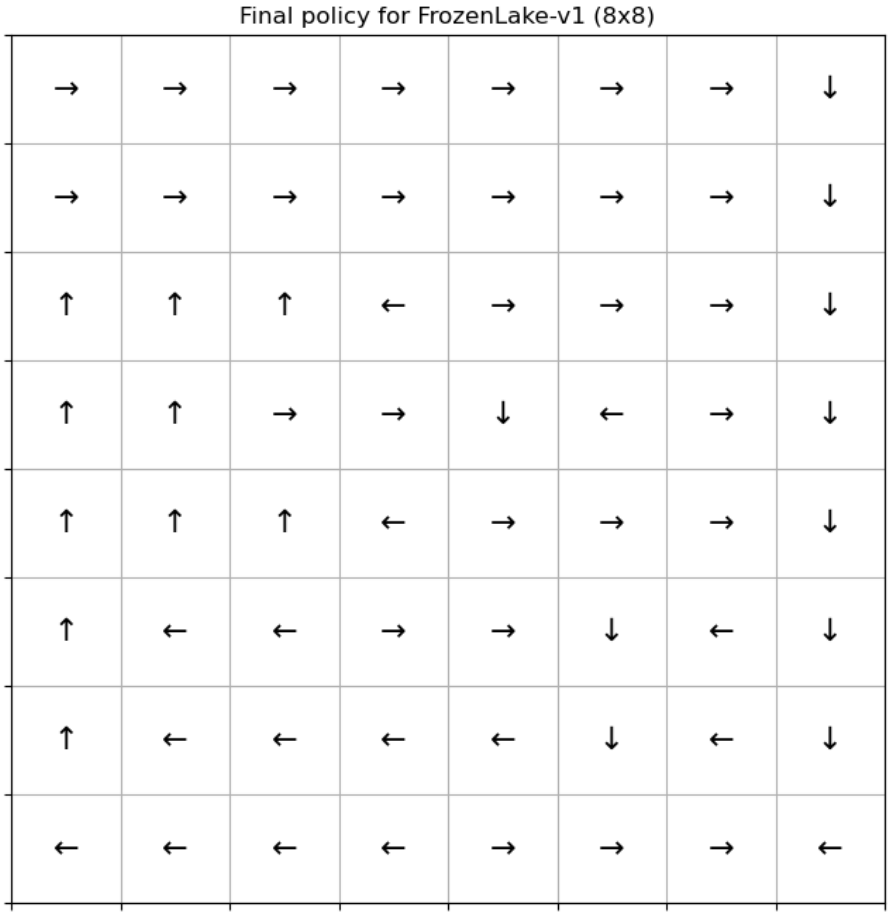
Monte Carlo First Visit with Exploring Starts

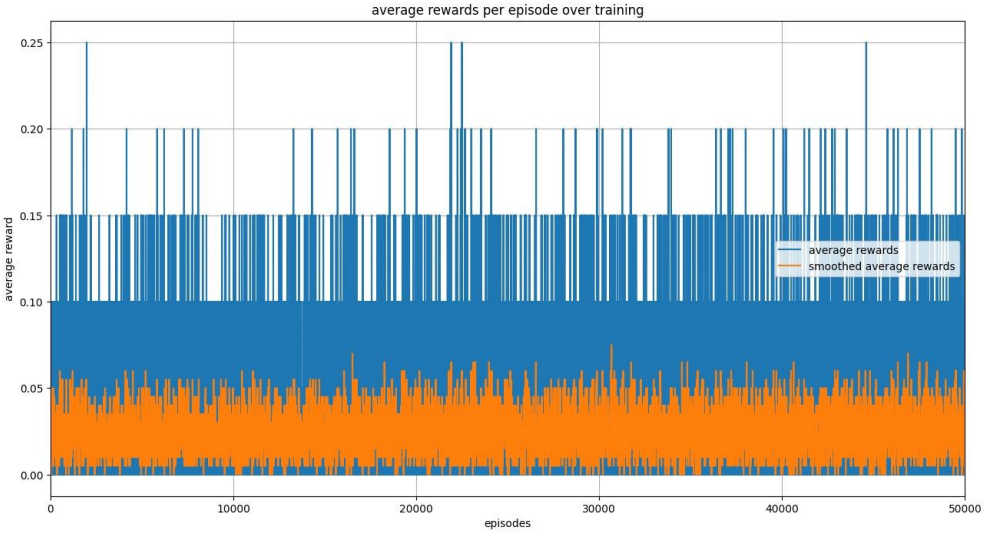
Slippery = False

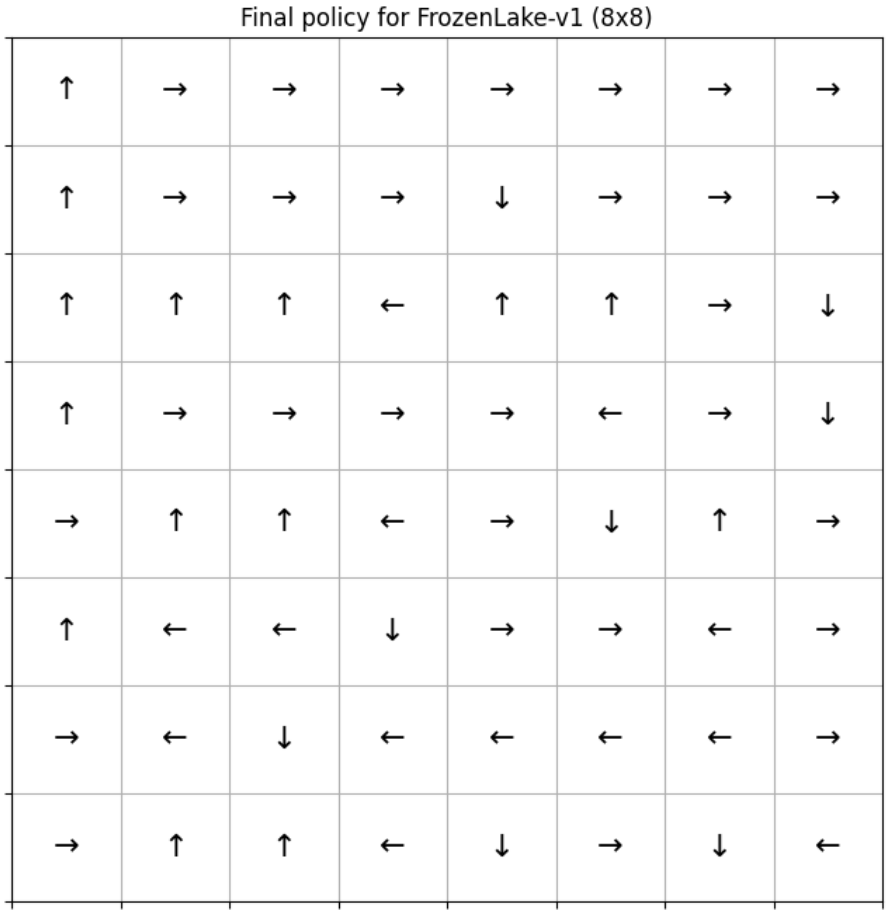




Monte Carlo First Visit with Exploring Starts

Slippery = True





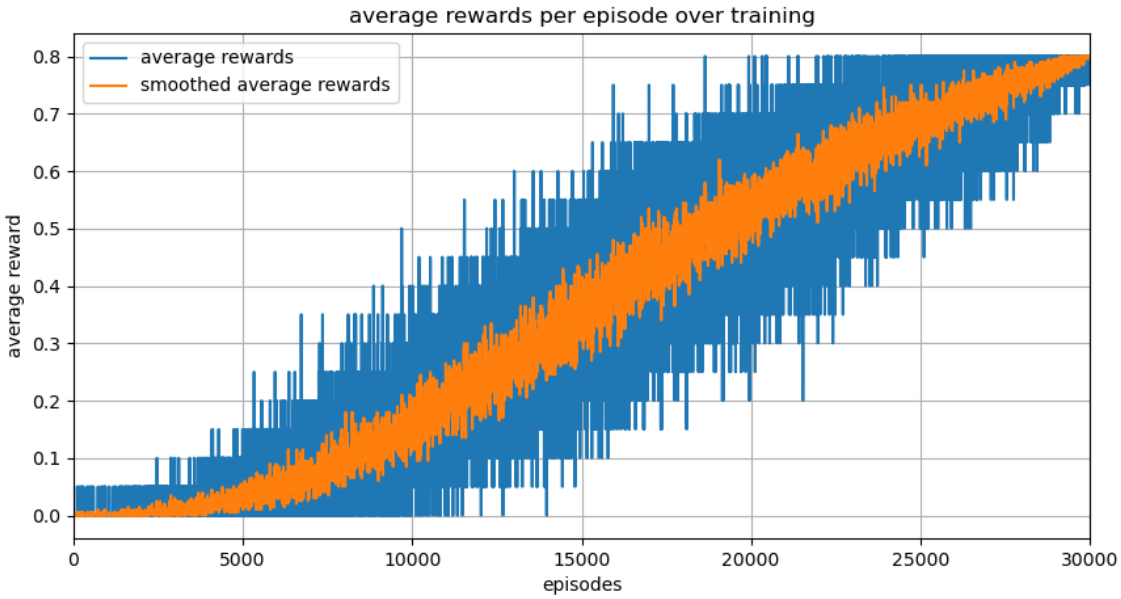
A diferencia de los métodos Temporal Diference que se ejecutan de manera online, Monte Carlo requiere conocer el episodio completo. Estos, al comenzar con una política sub-óptima, tienden a terminar con el agente cayendo en un hoyo o con el ambiente deteniéndose por alcanzar el máximo de movimientos permitidos (200 steps).

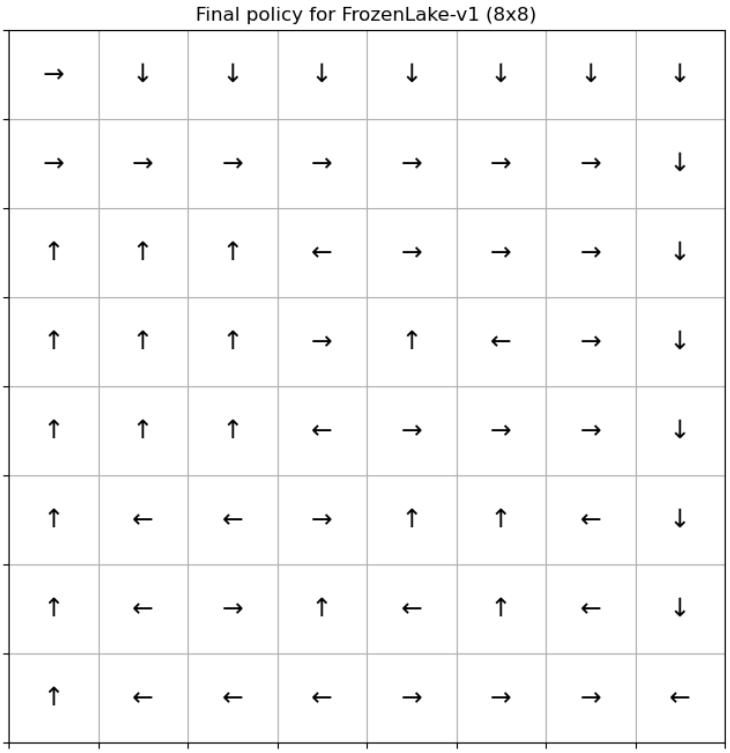
Dado que la recompensa se define como 1 cuando el agente alcanza la meta y 0 cuando cae en un hoyo y en todos los pasos intermedios, los únicos episodios que contribuyen a la actualización de la política son aquellos en los que el agente logra su objetivo, lo cual ocurre con baja frecuencia inicialmente. Así, en la mayoría de los episodios generados, el agente no realiza ningún aprendizaje.

Por otro lado, en los métodos Monte Carlo los valores de Q(s,a) sólo se actualiza con respecto a los retornos obtenidos en los episodios en los que se visitó el par estado-acción (s,a), a diferencia de los métodos Temporal Diference donde su valor también depende de los valores Q(s',a') de los pares (s',a') a los que se puede transicionar. Esto es, que en Monte Carlo los valores de Q no se propagan entre estados cercanos, lo que dificulta la convergencia.

Q-learning

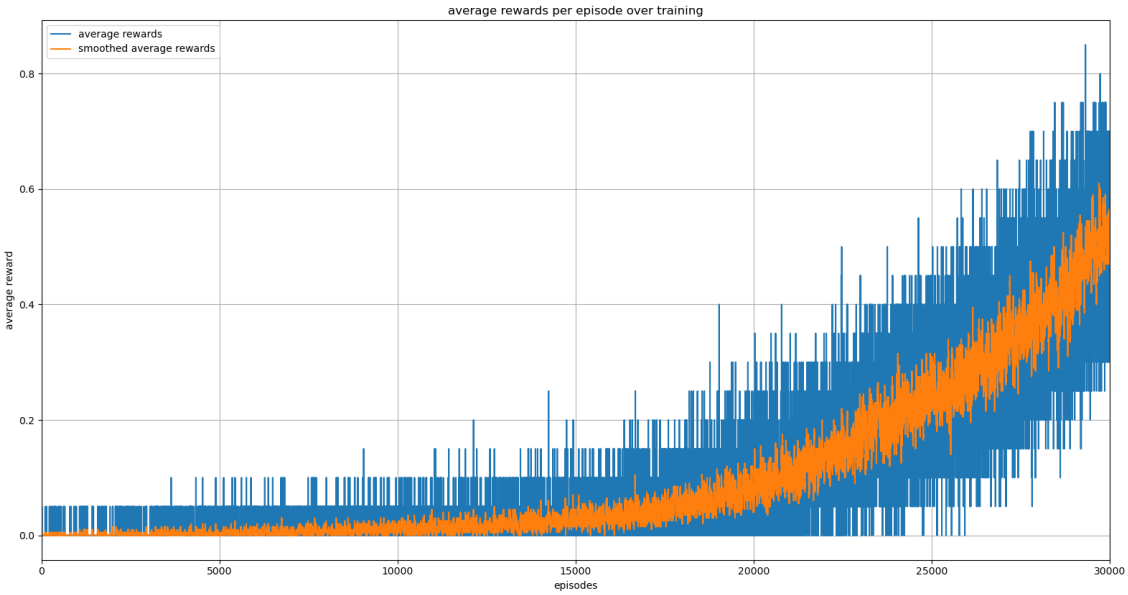
Slippery = False

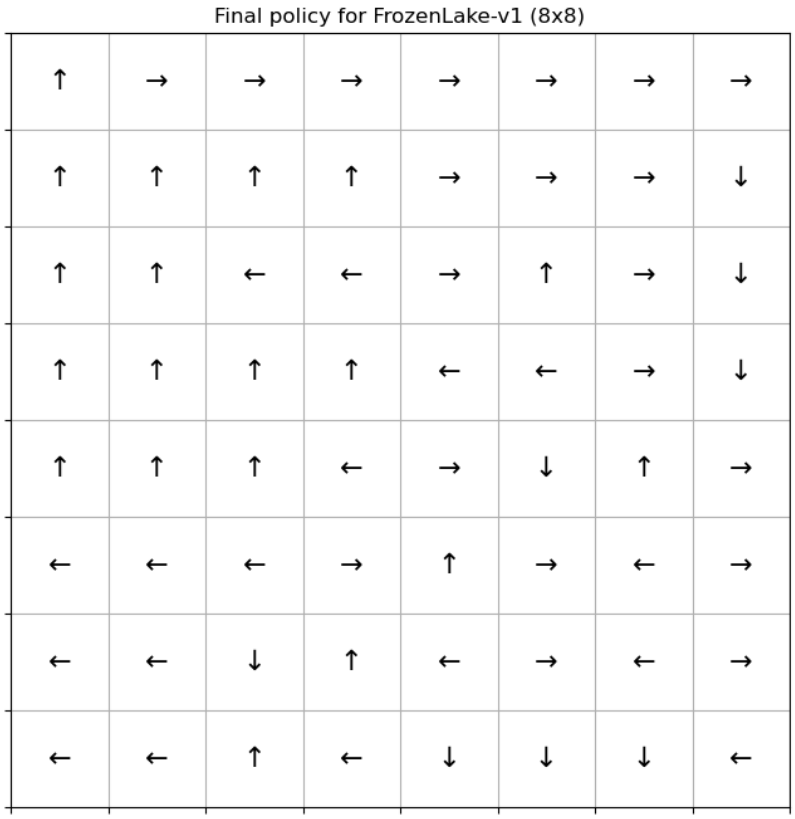




Q-learning

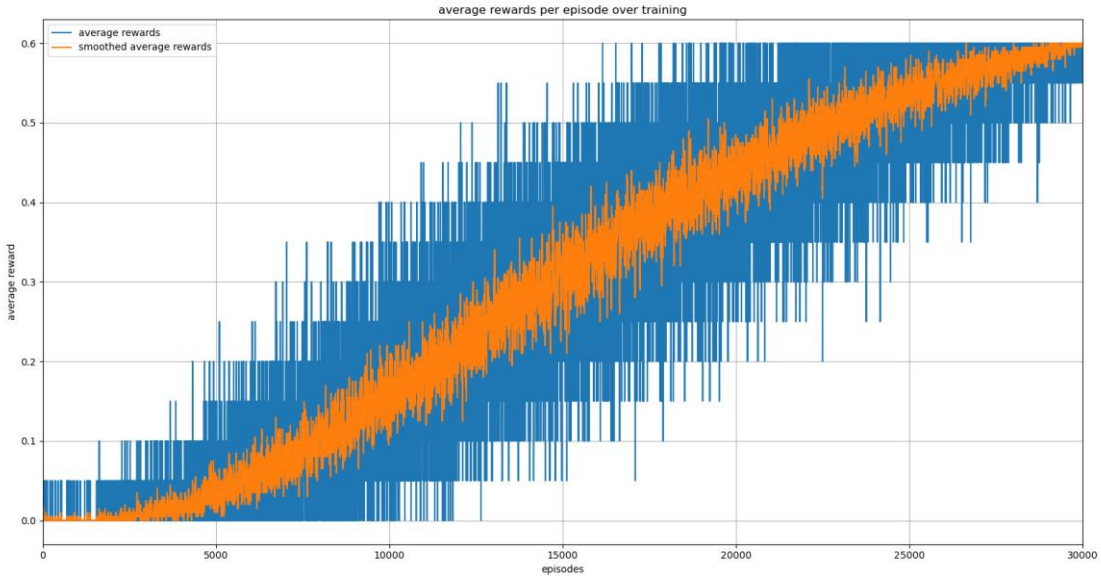
Slippery = True

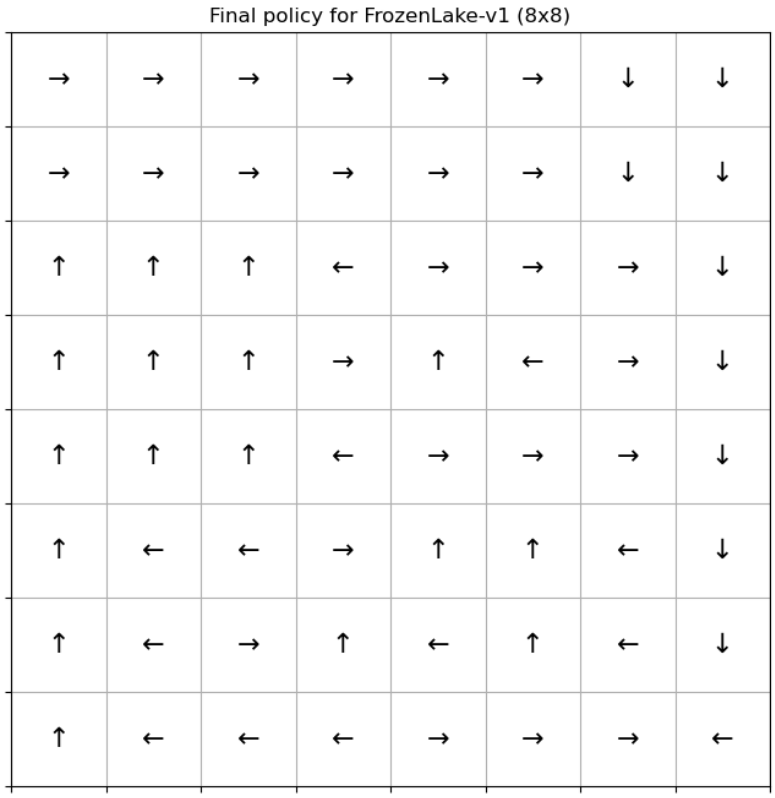




SARSA

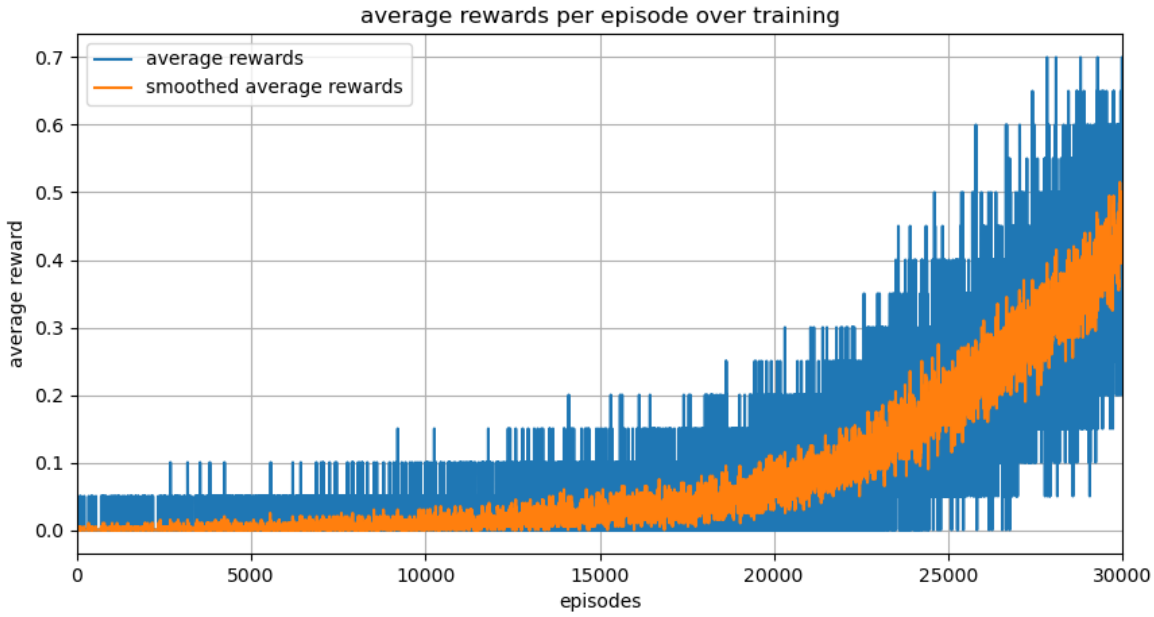
Slippery = False

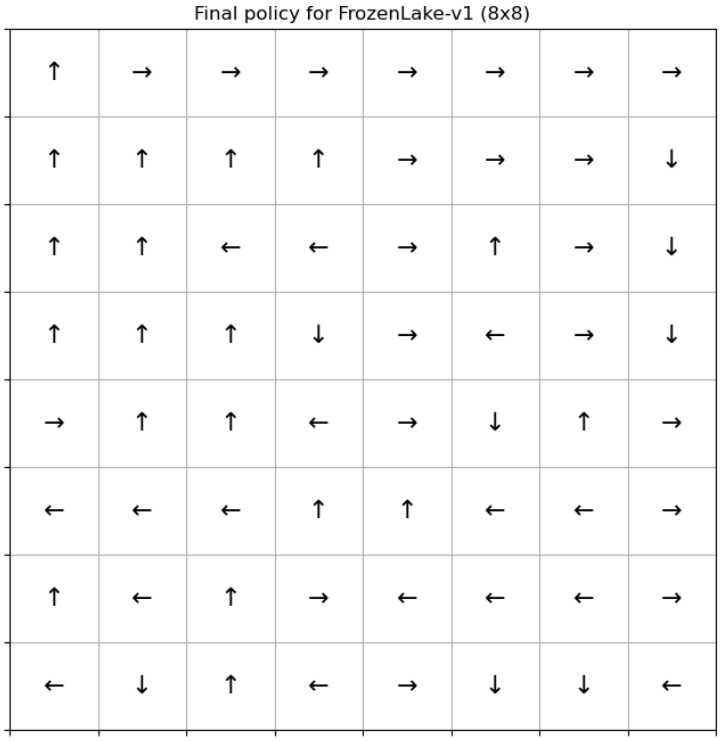




SARSA

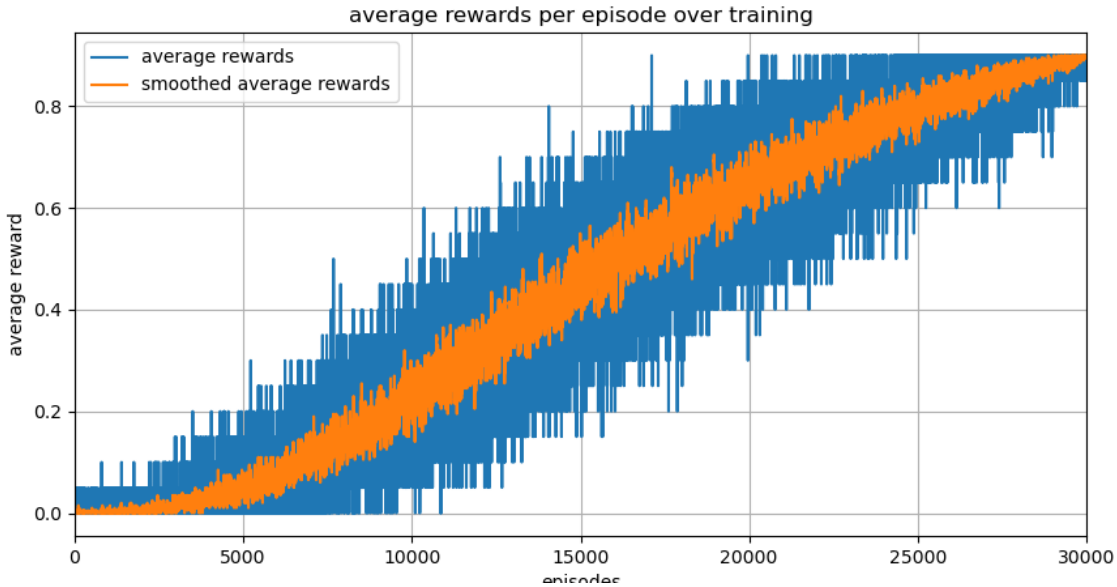
Slippery = True

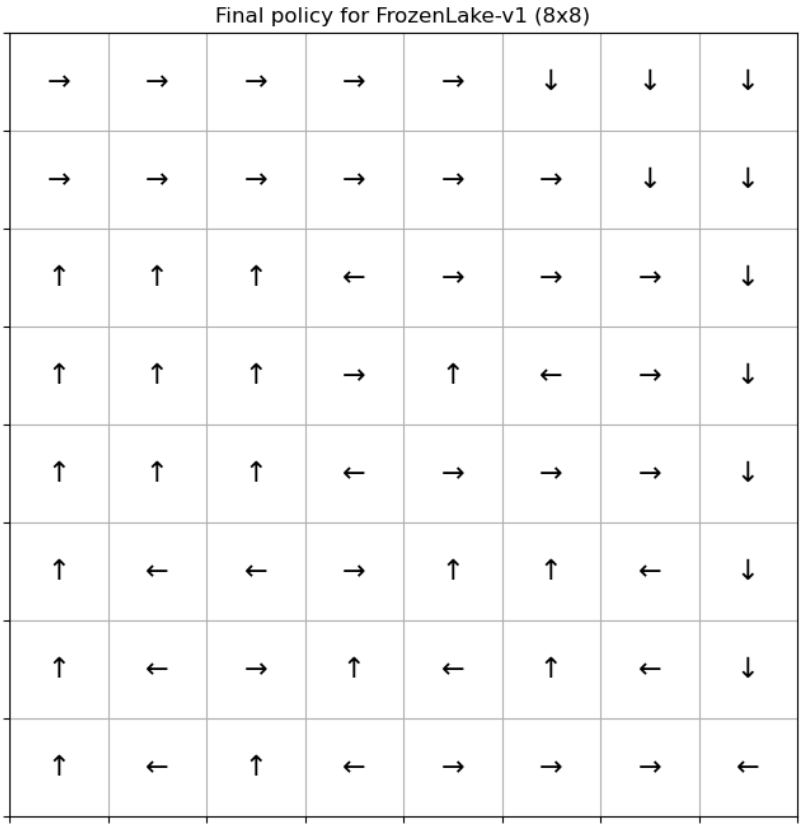




Expected-SARSA

Slippery = False





Expected-SARSA

Slippery = True