To overcome policy freezing a new MC approach was used to estimate the policy. This mainly involved using exploration to overcome the effects of policy freezing with two main methods being exploring starts and soft policies.

Instead of starting with a specific state and action, exploring starts result in episodes starting from different positions in the state space by choosing a random state and action such that all state-action pairs have a probability greater than 0. This approach can be seen in the second line in Figure 1.0. Using this approach causes the system to explore the whole space and when enough episodes are run, the algorithm will eventually realise that the random state-action pair selected is better than the existing policy, overwriting the suboptimal frozen policy. However, this requires randomly selecting the correct action for large state spaces with many suboptimalities, this approach is very inefficient and may take a long time to run.

Building on exploring starts, more exploration can be incorporated into the new MC approach by using soft policies. This is a type of policy that does not converge to a deterministic final function but instead retains an element of randomness when selecting actions. This allows the algorithm to explore and avoid freezing on suboptimal policies as it finds better ones to replace them with. An example of how this could be done is using ϵ-greedy soft policies. However, if this was the only policy used, the resulting optimal policy would have an ϵ-greedy mechanism embedded within it. This would be suboptimal as a portion of actions would be incorrect due to the random actions selected. Additionally, the rate of convergence of these algorithms is very slow, especially for long trajectories, as there are more and more possible paths leading to high variance and fewer samples available. To overcome this the algorithm could use smarter exploration strategies such as UCB which would favor less explored areas.

To overcome the embedded ϵ-greedy mechanism, the algorithm makes use of two policies. A deterministic target policy (on-policy) which is used after training and learning has been completed, and a behavioral policy (off-policy) which is used by the algorithm to learn the target policy. Over many episodes, the behavioral policy can be used to estimate the target policy as it will eventually converge. The code used to accomplish this is extracted from Sutton and Barton and is shown below in Figure X.

Text

Description automatically generated

One thing to note is that since the control policy is being estimated, the correction factor cannot be directly calculated. To convert the behavioral to the target policy, the code assumes that is equal to 1 and only updates when a deterministic action is taken.