Soft Actor-Critic for Discrete Action Settings (SAC-Discrete)  
Track the changes from theory to code.

Id = CHANGE0001  
i) It is now more efficient to have the soft Q-function output the Q-value of each possible action rather than simply the action provided as an input, i.e. our function moves from to . This was not possible be fore when there were infinitely many possible actions we could take.

Id = CHANGE0002  
ii) There is now no need for our policy to output the mean and covariance of our action distribution, instead it can directly output our action distribution. The policy therefore changes from to where now we are applying a softmax function in the final layer of the policy to ensure it outputs a valid probability distribution.

Id = CHANGE0003  
iii) Before, in order to minimise the soft Q-function cost (4) we had to plug in our sampled actions from the replay buffer to form a monte-carlo estimate of the soft state-value function (2). This was because estimating the soft state-value function in (2) involved taking an expectation over the action distribution. However, now, because our action set is discrete we can fully recover the action distribution and so there is no need to form a monte-carlo estimate and instead we can calculate the expectation directly. This change should reduce the variance involved in our estimate of the objective (4). This means that we change our soft state-value calculation equation from (2) to:

Id = CHANGE0004

iv) Similarly, we can make the same change to our calculation of the temperature loss to also reduce the variance of that estimate. The temperature objective changes from (9) to:

Id = CHANGE0005

v) Before, to minimise (6) we had to use the reparameterisation trick to allow gradients to pass through the expectations operator. However, now our policy outputs the exact action distribution we are able to calculcate the expectation directly. Therefore there is no need for the reparameterisation trick and the new objective for the policy changes from (8) to:

Combining all these changes, our algorithm for SAC with discrete actions (SAC-Discrete) is given by Algorithm I.