**Theoretical Answer**

The value of the advantage estimate reflects the contribution of the action beyond the current “average” of the state. Because the return can be change drastically between episodes, the REINFORCE algorithm gradient can suffer from high variance. This variance leads to unstable convergence. By subtracting the baseline, we can reduce the variability of the return and therefore the variance of the gradients.

1. The exaptation of the baseline during all episode transition according to policy can be express as the exaptation of the exaptation on each transition.
2. If (and only if) the baseline of state is independent with we can extract it from the expectation due to expectation linearity.
3. Expectation definition.
4. Log derivative.
5. According to Leibniz integral rule the gradient (by ) and the integral (by a) can be interchange.
6. the integral of a density function between and .
7. Gradient of constant equals to zero.

As mentioned in b prerequisite condition for this equation to be valid is that is independent with .