1.3 a) assuming $\ell(s,\pi)$ is 0-1 loss and

cost function C bounded in $[0,1]$

let $\mathbb{E}_{s\sim d_{\pi^*}}[\ell(s,\pi)] = \epsilon$

$\Delta C(s)$ $C(s,\pi^*)$ $C(s,\pi)$

* think of $\ell(s,\pi)$ as difference in cost suffered by optimal policy vs. learned policy given a state

* Then ΔC can be at most 1 b/c C is bounded in $[0,1]$, so the absolute difference b/w two values that are a result of $C(\cdot)$ is $|0-1|$ or $|1-0|$ at the most

* if $\ell(s,\pi)$ is the loss in cost suffered from not following π^* then $\ell(s,\pi)$ is also $= \Delta C(s)$

* since $s\sim d_{\pi^*}$, $\Delta C(s)$ can be expected to equal $\mathbb{E}_{s\sim d_{\pi^*}}[\ell(s,\pi)] = \epsilon$

First we have at a state by state basis:

$$\Delta C(s) = C(s,\pi) - C(s,\pi^*)$$

We expect these values to be

$$\underbrace{\mathbb{E}_{s\sim d_{\pi^*}}[\Delta C(s)]}_{\epsilon} = \mathbb{E}_{s\sim d_{\pi^*}}[C(s,\pi) - C(s,\pi^*)] = \mathbb{E}_{s\sim d_{\pi^*}}[C(s,\pi)] - \mathbb{E}_{s\sim d_{\pi^*}}[C(s,\pi^*)]$$

we rearrange to get

$$\mathbb{E}_{s\sim d_{\pi^*}}[C(s,\pi)] = \mathbb{E}_{s\sim d_{\pi^*}}[C(s,\pi^*)] + \epsilon$$

Over T time steps, the expected cost from following a policy accumulates to $J(\cdot)$

loss is accumulated at each step.

$$\hookrightarrow J(\pi) = J(\pi^*) + T\epsilon$$

since ϵ is at most $= 1$, at every step, we can accumulate a total loss of $T\cdot 1$ at most so:

$$J(\pi) \leq J(\pi^*) + T^2\epsilon$$