2.2.4 Reduction between control policies classes

We first show a reduction from a general history dependent policies to Randomized Markovian policies. The main observation is that the only influence on the cumulative cost is the expected instantaneous cost $\mathbb{E}[\mathtt{c}_\mathtt{t}(\mathtt{s}_\mathtt{t},\mathtt{a}_\mathtt{t})]$. Namely, let

$$\rho^\pi_\mathtt{t}(\mathtt{s},\mathtt{a}) = \Pr_{\mathtt{h}'_{\mathtt{t}-1}}[\mathtt{a}_\mathtt{t}=\mathtt{a},\mathtt{s}_\mathtt{t}=s] = \mathbb{E}_{\mathtt{h}'_{\mathtt{t}-1}}[\mathbb{I}[\mathtt{s}_\mathtt{t}=s,\mathtt{a}_\mathtt{t}=a]|\mathtt{h}'_{\mathtt{t}-1}],$$

where $\mathtt{h}'_{\mathtt{t}-1}=(\mathtt{s}_0,\mathtt{a}_0,\ldots,\mathtt{s}_{t-1},\mathtt{a}_{t-1})$ is the history of the first $\mathtt{t}-1$ time steps generated using π, and the probability and expectation are taken with respect to the randomness of the policy π. Now we can rewrite the expected cost to go as,

$$\mathbb{E}[\mathcal{C}^\pi(\mathtt{s}_0)] = \mathbb{E}[\sum_{\mathtt{t}=1}^{\mathtt{T}-1}\sum_{\mathtt{a}\in\mathcal{A}_\mathtt{t},\mathtt{s}\in\mathcal{S}_\mathtt{t}}\mathtt{c}_\mathtt{t}(\mathtt{s},\mathtt{a})\rho^\pi_\mathtt{t}(\mathtt{s},\mathtt{a})],$$

where $\mathcal{C}^\pi(\mathtt{s}_0)$ is the random variable of the cost when starting at state $\mathtt{s}_0$ and following policy π.

This implies that any two policies π and π' for which $\rho^\pi_\mathtt{t}(\mathtt{s},\mathtt{a})=\rho^{\pi'}_\mathtt{t}(\mathtt{s},\mathtt{a})$, for any time $\mathtt{t}$, state $\mathtt{s}$ and action $\mathtt{a}$, would have the same expected cumulative cost for any cost function, i.e., $\mathbb{E}[\mathcal{C}^\pi(\mathtt{s}_0)]=\mathbb{E}[\mathcal{C}^{\pi'}(\mathtt{s}_0)]$