Problems to be handed in

1 Prove Fact A and Fact B that were stated in the proof of Theorem 7.2 in the lecture on the two-armed bandit.

2 Consider a controlled Markov process with time-invariant controlled transition kernel $P_u(\mathrm{d}x'|x)$. Fix a nonnegative cost function $c : \mathcal{X} \times \mathcal{U} \to \mathbb{R}$ and a discount factor $\beta \in (0,1)$. For each $T \geq 0$, define the horizon-T discounted value function

$$V_T^\beta(x) := \inf_g \mathbf{E}^g \left[\sum_{t=0}^{T-1} \beta^t c(X_t, U_t) \middle| X_0 = x \right]. \tag{1}$$

Prove that the function computed after T steps of value iteration with zero initial condition is exactly the T-step β-discounted value function: for each $T > 0$,

$$V_T^\beta(x) = \mathbb{T}^\beta V_{T-1}(x), \qquad \forall x \in \mathcal{X} \tag{2}$$

with the initial condition $V_0^\beta(\cdot) = 0$, where $\mathbb{T}^\beta$ is the β-discounted dynamic programming operator that maps any measurable function $h : \mathcal{X} \to \mathbb{R}$ to

$$\mathbb{T}^\beta h(x) := \inf_{u \in \mathcal{U}} \left\{ c(x,u) + \beta \int_{\mathcal{X}} h(x') P_u(\mathrm{d}x'|x) \right\}. \tag{3}$$

3 Consider the setting of Problem 2. Any measurable mapping $\phi : \mathcal{X} \to \mathcal{U}$ defines a stationary Markov policy $\phi^\infty = (\phi, \phi, \ldots)$, i.e., the action at each time t is given by $U_t = \phi(X_t)$. Given a discount factor $\beta \in (0,1)$, define the discounted cost of ϕ^∞:

$$J^\beta(x; \phi^\infty) := \mathbf{E}^{\phi^\infty} \left[\sum_{t=0}^{\infty} \beta^t c(X_t, \phi(X_t)) \middle| X_0 = x \right].$$

(i) Prove that, for any $x \in \mathcal{X}$,

$$J^\beta(x; \phi^\infty) = c(x, \phi(x)) + \beta \int_{\mathcal{X}} J^\beta(x'; \phi^\infty) P_{\phi(x)}(\mathrm{d}x'|x). \tag{4}$$

(ii) Suppose the cost c is nonnegative and bounded. Prove that the fixed-point equation

$$w(x) = c(x, \phi(x)) + \beta \int_{\mathcal{X}} w(x') P_{\phi(x)}(\mathrm{d}x'|x) \tag{5}$$

has a unique bounded solution, which is equal to $J^\beta(x; \phi^\infty)$.

Hint: Use the contraction mapping principle.

4 Consider the setting of Problems 2 and 3, but now with *finite* state and action spaces. Prove that the policy iteration algorithm terminates after finitely many iterations and produces an optimal policy.

Hint: Show that the policy computed at each iteration n is better than the policy computed at iteration $n-1$.