Exercise 7.20

(a) This problem is exactly like Example 7.5.3, with the same Bellman's equation:

$$h^*(i) = \min\{-b_i - \lambda^*(T_i + \frac{1}{r}) + \sum_{j=1}^{n} p_j h^*(j), -\lambda^* \frac{1}{r} + \sum_{j=1}^{n} p_j h^*(j)\}$$

which results in the same optimal policy: accept offer (b_i, T_i) if $\frac{b_i}{T_i} \geq -\lambda^*$, where $-\lambda^*$ is the optimal average benefit per unit time.

(b) Let us assume without loss of generality that $\frac{b_1}{T_1} \leq \frac{b_2}{T_2}$. There are four possible policies: reject both types of jobs, accept only type 1 jobs, accept only type 2 jobs, or accept both types of jobs. If we reject both types of jobs, then the average income per unit time, λ, is 0. This policy cannot be optimal because both b_1 and b_2 are positive. If accepting type 1 jobs is optimal, then it must also be optimal to accept type 2 jobs because we have assumed that $\frac{b_1}{T_1} \leq \frac{b_2}{T_2}$. Therefore accepting only type 1 jobs is not optimal. Both remaining possible policies accept type 2 jobs, so we know it is optimal to accept type 2 jobs. We now find the average benefit per unit time for both remaining policies.

Reject type 1 jobs, Accept type 2 jobs:

$$\begin{aligned}
h(1) &= -\lambda\frac{1}{r} + p_1 h(1) + p_2 h(2) \\
h(2) &= -b_2 - \lambda(T_2 + \frac{1}{r}) + p_1 h(1) + p_2 h(2) = 0 \\
\Rightarrow -\lambda &= \frac{r p_2 b_2}{1 + r p_2 T_2}
\end{aligned}$$

Accept type 1 jobs, Accept type 2 jobs:

$$\begin{aligned}
h(1) &= -b_1 - \lambda(T_1 + \frac{1}{r}) + p_1 h(1) + p_2 h(2) \\
h(2) &= -b_2 - \lambda(T_2 + \frac{1}{r}) + p_1 h(1) + p_2 h(2) = 0 \\
\Rightarrow -\lambda &= \frac{r p_1 b_1 + r p_2 b_2}{1 + r p_1 T_1 + r p_2 T_2}
\end{aligned}$$

The latter policy is optimal if and only if:

$$\begin{aligned}
\frac{r p_2 b_2}{1 + r p_2 T_2} &\leq \frac{r p_1 b_1 + r p_2 b_2}{1 + r p_1 T_1 + r p_2 T_2} \\
\Leftrightarrow \frac{b_2}{T_2} &\leq (1 + \frac{1}{r p_2 T_2})\frac{b_1}{T_1}
\end{aligned}$$

The optimal policy accepts offer 2 for any values of $r, p_1, p_2, b_1, b_2, T_1, T_2$ and accepts offer 1 if and only if $\frac{b_2}{T_2} \leq (1 + \frac{1}{r p_2 T_2})\frac{b_1}{T_1}$.

Therefore, the optimal average benefit per unit time is equal to $\frac{r p_2 b_2}{1 + r p_2 T_2}$ if $(1 + \frac{1}{r p_2 T_2})\frac{b_1}{T_1} < \frac{b_2}{T_2}$ and is equal to $\frac{r p_1 b_1 + r p_2 b_2}{1 + r p_1 T_1 + r p_2 T_2}$ otherwise.

(c) Recall the following about exponential random variables. If X is exponentially distributed with parameter x and Y is exponentially distributed with parameter y, then $\min(X, Y)$ is exponentially distributed with parameter $x + y$ and $P(X \leq Y) = \frac{x}{x+y}$.

Because the system can process more than one job simultaneously, we now augment the state with the number and types of jobs currently being processed. For the case where $m = 2$, the state becomes (i, s), where i is the current job offer being considered and $s = 0$ if no jobs are currently being processed and $s = l$ if a job of type l is currently being processed.

If the system is currently processing less than m jobs, then the state changes when the next job offer arrives. If the system is currently processing m jobs, then the state changes after one job completes and when the next job offer arrives. In either case, the new state depends on which jobs have already finished by the time the next job offer arrives.

We have the following Bellman's equation for $m = 2$.

$$
\begin{aligned}
h^*(i,0) = \min\{&-b_i - \lambda^* \frac{1}{r} + \sum_{j=1}^{n} p_j \Big(\frac{r}{T_i^{-1} + r} h^*(j,i) + \frac{T_i^{-1}}{T_i^{-1} + r} h^*(j,0)\Big), \\
&- \lambda^* \frac{1}{r} + \sum_{j=1}^{n} p_j h^*(j,0)\} \\
h^*(i,l) = \min\{&-b_i - \lambda^* \Big(\frac{1}{T_i^{-1} + T_l^{-1}} + \frac{1}{r}\Big) \\
&+ \sum_{j=1}^{n} p_j \Big(\frac{T_i^{-1}}{T_i^{-1} + T_l^{-1}} \Big(\frac{r}{T_l^{-1} + r} h^*(j,l) + \frac{T_l^{-1}}{T_l^{-1} + r} h^*(j,0)\Big) \\
&+ \frac{T_l^{-1}}{T_i^{-1} + T_l^{-1}} \Big(\frac{r}{T_i^{-1} + r} h^*(j,i) + \frac{T_i^{-1}}{T_i^{-1} + r} h^*(j,0)\Big)\Big), \\
&- \lambda^* \frac{1}{r} + \sum_{j=1}^{n} p_j \Big(\frac{r}{T_l^{-1} + r} h^*(j,l) + \frac{T_l^{-1}}{T_l^{-1} + r} h^*(j,0)\Big)\}
\end{aligned}
$$

The exponential distribution assumption is necessary because an exponential random variable is memoryless. More specifically, at any point in time, the time until the next job offer (or until a current job completes) is still exponentially distributed with the same parameter no matter how much time has passed since the last job offer (or since the job first began).

Exercise 7.22

(a) This is a stochastic shortest path problem with state equal to the number of treasures not yet found. The termination state is state 0, and we assume that when the state moves to 0, the hunter decides to stop searching. When the hunter decides to search, the state moves from i to $i - m$ with probability $p(m \mid i)$.

(b) Bellman's equation is

$$J(i) = \max\left[0, r(i) - c + \sum_{m=0}^{i} p(m \mid i) J(i-m)\right], \qquad i = 1, \ldots, n,$$

$$J(0) = 0.$$

It has a unique solution because under our assumption $p(0 \mid i) < 1$ for all i, the termination state is reached with probability 1 under all policies, and the assumptions required by the results of Chapter 7 are satisfied.

(c) The policy that never searches has value function $J_{\mu^0}(i) = 0$ for all i. The policy μ^1 subsequently produced by policy iteration is the one that searches at a state i if and only if $r(i) > c$, and has corresponding value function

$$J_{\mu^1}(i) = \begin{cases} 0 & \text{if } r(i) \le c, \\ r(i) - c + \sum_{m=0}^{i} p(m \mid i) J_{\mu^1}(i-m) & \text{if } r(i) > c, \end{cases}$$

Its value function is nonnegative for all i, which can be shown by induction using the value iteration

$$J_{k+1}(i) = \begin{cases} 0 & \text{if } r(i) \le c, \\ r(i) - c + \sum_{m=0}^{i} p(m \mid i) J_k(i-m) & \text{if } r(i) > c, \end{cases}$$

Assuming $J_k(i) \ge 0$ for all i, we have $J_{k+1}(i) \ge 0$ for all i. Starting from any $J_0(i) \ge 0$ for all i, we then have $J_k(i) \ge 0$ for all i, k, meaning its limit is also nonnegative, i.e. $J_{\mu^1}(i) \ge 0$ for all i.

The next policy is obtained from the minimization

$$\mu^2(i) = \arg\max\left[0, r(i) - c + \sum_{m=0}^{i} p(m \mid i) J_{\mu^1}(i-m)\right], \qquad i = 1, \ldots, n.$$

For i such that $r(i) \le c$, we have $r(j) \le c$ for all $j < i$ because $r(i)$ is monotonically nondecreasing in i. Therefore for i such that $r(i) \le c$, we have $J_{\mu^1}(i-m) = 0$ for all $m \ge 0$, so

$$0 \ge r(i) - c + \sum_{m=0}^{i} p(m \mid i) J_{\mu^1}(i-m),$$

and $\mu^2(i) =$ *stop searching*.

For i such that $r(i) > c$, $\mu^2(i) =$ *search* since $J_{\mu^1}(i) \ge 0$ for all i, so that

$$0 < r(i) - c + \sum_{m=0}^{i} p(m \mid i) J_{\mu^1}(i-m).$$

Thus, μ^2 is the same as μ^1, so it is optimal.

Exercise 7.23

Let the state be the current set of wins.

(a)

$$J^*(0) = \min_i \{c_i + p_i J^*(i) + (1 - p_i) J^*(0)\}$$

$$J^*(i) = \min_{j \neq i} \{c_j + p_j J^*(i,j) + (1 - p_j) J^*(0)\}$$

$$J^*(i,j) = c_k - p_k m + (1 - p_k) J^*(0) \qquad k \neq i, j$$

(b) Let J represent the cost-to-go for the stationary policy ijk. Then J satisfies the following set of equations:

$$J(0) = c_i + p_i J(i) + (1 - p_i) J(0)$$

$$J(i) = c_j + p_j J(i,j) + (1 - p_j) J(0)$$

$$J(i,j) = c_k - p_k m + (1 - p_k) J(0)$$

Solving the above equations for $J(0)$, the expected cost of policy ijk, we have:

$$J(0) = \frac{c_i + p_i c_j + p_i p_j c_k - p_i p_j p_k m}{p_i p_j p_k}$$

(c)

$$\lambda^* + h^*(0) = \min_i \{c_i + p_i h^*(i) + (1 - p_i) h^*(0)\}$$

$$\lambda^* + h^*(i) = \min_{j \neq i} \{c_j + p_j h^*(i,j) + (1 - p_j) h^*(0)\}$$

$$\lambda^* + h^*(i,j) = c_k + p_k(-m + h^*(0)) + (1 - p_k) h^*(0) \qquad k \neq i, j$$

(d) Let λ represent the average cost-per-stage for the stationary policy ijk. Then λ satisfies the following set of equations:

$$\lambda + h(0) = c_i + p_i h(i) + (1 - p_i) h(0)$$

$$\lambda + h(i) = c_j + p_j h(i,j) + (1 - p_j) h(0)$$

$$\lambda + h(i,j) = c_k + p_k(-m + h(0)) + (1 - p_k) h(0) \qquad k \neq i, j$$

Exercise 7.24

(a) Consider the single-policy average cost problem involving the Markov chain with transition probabilities p_{ij}, and cost at state i equal to y^i, $i = 1, \dots, s$. Assume without loss of generality that state 1 is a recurrent state. Since there is only one recurrent class, assumption 7.4.1 is satisfied. Thus $\overline{y}$ is the average cost per stage (common for all states) and solves uniquely, together with scalars $h(i)$, $i = 1, \dots, s$, Bellman's equation, which is

$$\overline{y} + h(i) = y^j + \sum_{j=1}^{s} p_{ij} h(j), \qquad i = 1, \dots, s,$$

$$h(1) = 0.$$

(b) The states of the problem are of two types:

(1) Pairs (current offer w_k, current value of y_k). These are the $m \cdot s$ states where the house is still being rented, and there is a choice whether to accept or reject the current offer. The reward in these states is R if the offer is rejected, and is $y_k w_k$ if the offer is accepted.

(2) Pairs (price at which the house has been sold w_k, current value of y_k). These are the $m \cdot s$ states that occur after the house has been sold. The value of w_k stays constant when transitioning from these states, and the value of y_k evolves according to the transition probabilities p_{ij}. The reward in these states is $y_k w_k$.

There are two types of stationary policies: (i) The policy that never sells, which is optimal if and only if $R \geq w^i \overline{y}$ for all possible values of the offers w^i. (ii) A policy that sells for some pairs (current offer w_k, current value of y_k) but not for others. The average cost per stage then is $w\overline{y}$ where w is the sale price. Thus the value of y_k does not matter in accepting the current offer. Clearly, among these policies, the optimal one sells when the price has reached its maximum (which will happen with probability 1) and the corresponding average cost per stage is $\overline{wy}$. Thus, even if the maximum offer is extremely rare, it is optimal to wait for it.

The average cost formulation is not satisfactory because alot of counterintuitive transient behavior would be optimal. For example, a (nonstationary) policy that rejects all offers up to a certain (arbitrarily large) time period and then waits to accept the maximum offer is also optimal. So the average cost formulation does not provide any penalty for delaying a sale decision, even when the current value of y_k is favorable and the current offer is high but not maximal.

(c) Let the state be (w_k, y_k, r), where $r = 0$ means the house is being rented and $r = 1$ means the house has been sold. We then have the following Bellman's equation for $i = 1, 2, \ldots, m$, $j = 1, 2, \ldots, s$:

$$J^*(i,j,0) = \max\{\underbrace{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_{jl} J^*(k,l,0)}_{reject},\ \underbrace{w^i y^j + \alpha \sum_{l=1}^{s} p_{jl} J^*(i,l,1)}_{sell}\},$$

$$J^*(i,j,1) = w^i y^j + \alpha \sum_{l=1}^{s} p_{jl} J^*(i,l,1).$$

Let $\overline{y}(j)$ be the unique solution to $\overline{y}(j) = y^j + \alpha \sum_{l=1}^{s} p_{jl} \overline{y}(l)$ for $j = 1, 2, \ldots, s$ (which exists because the y^j are bounded for all j and there are s equations with s unknowns). Let $J(i,j,1) = w^i \overline{y}(j)$, and plug it into the righthandside of Bellman's equation for $r = 1$:

$$\begin{aligned}
J^*(i,j,1) &= w^i y^j + \alpha \sum_{l=1}^{s} p_{jl} J^*(i,l,1) \\
&= w^i y^j + \alpha \sum_{l=1}^{s} p_{jl} w^i \overline{y}(l) \\
&= w^i (y^j + \alpha \sum_{l=1}^{s} p_{jl} \overline{y}(l)) \\
&= w^i \overline{y}(j)
\end{aligned}$$

Because J satisfies Bellman's equation, J is the optimal cost-to-go, so $J^*(i,j,1) = w^i \overline{y}(j)$. Plugging this into the righthandside of Bellman's equation for $r = 0$:

$$\begin{aligned}
J^*(i,j,0) &= \max\{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_{jl} J^*(k,l,0),\ w^i y^j + \alpha \sum_{l=1}^{s} p_{jl} J^*(i,l,1)\} \\
&= \max\{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_{jl} J^*(k,l,0),\ w^i y^j + \alpha \sum_{l=1}^{s} p_{jl} w^i \overline{y}(l)\} \\
&= \max\{\underbrace{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_{jl} J^*(k,l,0)}_{reject},\ \underbrace{w^i (y^j + \alpha \sum_{l=1}^{s} p_{jl} \overline{y}(l))}_{sell}\}
\end{aligned}$$

The optimal policy is therefore to accept an offer if and only if

$$w^i \geq \frac{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_{jl} J^*(k,l,0)}{y^j + \alpha \sum_{l=1}^{s} p_{jl} \overline{y}(l)} = T(j).$$

(d) We assume that $p_{ij} = p_j$ for all i and $y^1 < y^2 < \ldots < y^s$.

i. We have the following expression for the threshold:

$$T(j) = \frac{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_l J^*(k, l, 0)}{y^j + \alpha \sum_{l=1}^{s} p_l \overline{y}(l)}$$

Notice that in the expression above for $T(j)$, both the numerator and the second term in the denominator are constant with respect to j. Because y^j is monotonically increasing in j, we have $T(j)$ is monotonically decreasing in j. So the higher the yield y_k, the lower the threshold for w_k, and the more likely we are to accept an offer.

ii. Consider any base policy. Denote its cost-to-go when renting as $J_{base}(i, j, 0)$. The corresponding rollout policy accepts at state (w^i, y^j) if and only if

$$w^i \geq \frac{R + \alpha \sum_{k=1}^{m} \sum_{l=1}^{s} q^k p_l J_{base}(k, l, 0)}{y^j + \alpha \sum_{l=1}^{s} p_l \overline{y}(l)}.$$

Once again, the righthandside of the above inequality is monotonically decreasing in j. Therefore, policy iteration, when started with a policy of monotonically decreasing thresholds (or any policy in general), generates a sequence of policies with monotonically decreasing thresholds.

Exercise 7.25

(a) Let the state be the current consecutive number of days parked on the street. Because there exists m such that $p_m T > G$, there exists n, where n is the smallest integer j such that $p_j T > G$. Because we would never park on the street if its expected cost were more than parking in the garage, we may define the state space as $\{0, 1, \ldots, n\}$, which is finite, and still find the optimal policy for the problem with infinite state space $\{0, 1, \ldots\}$.

$$J^*(i) = \min[G + \alpha J^*(0), p_{i+1} T + \alpha J^*(i+1)] \qquad i = 0, 1, \ldots, n-1$$

$$J^*(n) = G + \alpha J^*(0)$$

(b) Use VI to show $J^*(i)$ is monotonically nondecreasing in i, and therefore the optimal policy has a threshold form: park on the street if $i \leq i^*$ and park in the garage if $i > i^*$. Assume $J_k(i)$ is monotonically nondecreasing in i. We find J_{k+1} according to:

$$J_{k+1}(i) = \begin{cases} \min[G + \alpha J_k(0), p_{i+1} T + \alpha J_k(i+1)] & i = 0, 1, \ldots, n-1 \\ G + \alpha J_k(0) & i = n \end{cases}$$

Because p_{i+1} and $J_{k+1}(i)$ are monotonically nondecreasing in i, we know the righthand-side of the minimization is nondecreasing in i, meaning $J_{k+1}(i)$ is nondecreasing for $i < n$. We also have

$$J_{k+1}(n-1) = \min[G + \alpha J_k(0), p_n T + \alpha J_k(n)] \leq G + \alpha J_k(0) = J_{k+1}(n)$$

Therefore, $J_{k+1}(i)$ is monotonically nondecreasing in i for all i. Starting from any $J_0(i)$ that is monotonically nondecreasing, we have by induction that $J_k(i)$ monotonically nondecreasing in i for all k, meaning $J^*(i) = \lim_{k\to\infty} J_k(i)$ is monotonically nondecreasing in i.

(c) We show if we start policy iteration with a threshold policy, then we generate a sequence of threshold policies, meaning policy iteration takes no more than n, the number of possible thresholds, iterations.

Consider the policy that parks on the street at state i if and only if $i \leq m$, where $m \leq n-1$. Then we have the following corresponding cost-to-go:

$$J_m(i) = \begin{cases} p_{i+1}T + \alpha J_m(i+1), & i \leq m \\ G + \alpha J_m(0), & i > m \end{cases}$$

Letting $G + \alpha J_m(0) = c$, we have the following rollout policy for $i = 0, 1, \ldots, n-1$:

$$\mu_r(i) = argmin[c, p_{i+1}T + \alpha J_m(i+1)]$$

Consider the case where $J_m(m) \leq c$:

We show that $J_m(i)$ is monotonically nondecreasing in i. For $i > m$, we have $J_m(i) = c$ which is constant. For $i \leq m$, recall for this case we have $J_m(m) \leq c = J_m(m+1)$, which provides a base case for showing that $J_m(i) \leq J_m(i+1)$ for $i \leq m$. Assuming for induction that $J_m(i) \leq J_m(i+1)$, we have

$$J_m(i-1) = p_i T + \alpha J_m(i) \leq p_{i+1}T + \alpha J_m(i+1) = J_m(i).$$

Therefore $J_m(i)$ is monotonically nondecreasing in i for all i, meaning the right-handside of the rollout policy minimization is monotonically nondecreasing in i, and the rollout policy has a threshold form.

Consider the case where $J_m(m) > c$:

Re-expressing the rollout policy using Bellman's equation for $J_m(i)$, we have:

$$\begin{aligned} \mu_r(i) &= argmin[c, p_{i+1}T + \alpha J_m(i+1)] \\ &= \begin{cases} argmin[c, p_{i+1}T + \alpha c] & i \geq m \\ argmin[c, J_m(i)] & i < m \end{cases} \end{aligned}$$

Assume for contradiction there exists $\bar{i}$ such that the rollout policy parks in the garage at $\bar{i}$ and on the street at $\bar{i}+1$.

If $\bar{i} \geq m$, we have $c \leq p_{\bar{i}+1}T + \alpha c$ and $c > p_{\bar{i}+2}T + \alpha c$, which contradicts that $p_{\bar{i}+1} \leq p_{\bar{i}+2}$.

We know $\bar{i} \neq m-1$ because we have assumed $J_m(m) > c$, which means the rollout policy parks in the garage at m.

If $\bar{i} < m-1$, we have $c < J_m(\bar{i})$ and $c \geq J_m(\bar{i}+1)$, meaning $J_m(\bar{i}) > J_m(\bar{i}+1)$. Because $c \geq J_m(\bar{i}+1)$ and $c < J_m(m)$, there exists j such that $\bar{i} < j < m$ and $J_m(j) \leq J_m(j+1)$. By the induction proof in the previous case, we then know $J_m(i) \leq J_m(i+1)$ for $i \leq j$, which contradicts $J_m(\bar{i}) > J_m(\bar{i}+1)$.

Therefore, no such $\bar{i}$ exists, and if the rollout policy parks in the garage at i, then it parks in the garage at all states $j > i$, which is a threshold policy.

(d) As for the discounted cost version, we may consider only states $\{0, 1, \ldots, n\}$. State 0 is recurrent under all policies, so we may apply Bellman's equation:

$$\lambda^* + h^*(i) = \min[G + h^*(0), p_{i+1}T + h^*(i+1)], \qquad i = 0, 1, \ldots, n-1,$$

$$\lambda^* + h^*(n) = G + h^*(0),$$

$$h^*(0) = 0.$$

Exercise 7.26

(a) Let the set of possible offers be $\{x_i | i = 1, \ldots, n\} \cup \{t, r\}$, where being in a state has the following meaning:

$$x_i\text{: receiving offer } s_i; \quad t\text{: overtaken}; \quad r\text{: retired.}$$

Define the control space by $\{A(\text{accept offer}), I(\text{invest}), R(\text{retire})\}$. The state r is absorbing, and for the other states, the set of admissible controls is

$$U(x_i) = \{A, R\}, \qquad U(t) = \{I, R\}.$$

Define the corresponding transition probabilities, and per-stage costs as described in the problem. Bellman's equation for $\alpha < 1$ is

$$J^*(x_i) = \max\left[s_i, \alpha\left((1-\beta)\sum_{j=1}^{n} p_j J^*(x_j) + \beta J^*(t)\right)\right], \tag{1}$$

$$J^*(t) = \max\left[0, -v + \alpha\left(\gamma\sum_{j=1}^{n} p_j J^*(x_j) + (1-\gamma)J^*(t)\right)\right]. \tag{2}$$

(b) From Eq. (1), we see that since the second term in the minimization of the right-hand side does not depend on the current state x_i, the optimal policy is a threshold policy at states x_i, $i = 1, \ldots, n$. From Eq. (2), we see that once the inventor has been overtaken, it is either optimal to retire immediately, or to keep investing until his mouse trap is improved.

(c) Even without discounting, all policies have finite total cost, except the one that never sells and always invests, which incurs infinite total cost. Thus, we see that under the average cost criterion, all of the finite total cost policies will be optimal. Thus, the average cost criterion does not discriminate enough between different policies and makes no sense.

(d) Since there is no discounting and $v = 0$, there is no penalty for waiting as long as necessary until the maximum possible offer is received, which is going to happen with probability 1. So a stochastic shortest path formulation makes limited sense, since it excludes from consideration all offers except the maximum possible, no matter how unlikely this maximum offer is.