Assignment - Week I.

Corentin Lepla
Student number: 2898108

Milan Peter
Student number: 2868506

1 State Space and Action Space

We model time in discrete periods $t = 1, \dots, T$ with a horizon $T = 150$.

Our **state** x_t is the on-hand inventory at the start of each period t:

$$x_t \in \mathcal{X} = \{0, 1, \dots, X_{\max}\}.$$

This state is sufficient as it captures all necessary information from the past (it has the Markov property). To ensure our algorithm is efficient and runs fast, we must set a finite cap on the state space. We choose $X_{\max} = 25$, which is a safe upper bound given the initial inventory is $x_1 = 5$ and the maximum possible net inventory increase per period is 1 (an order arrives $A_t = 1$ and demand is $D_t = 0$).

Our **action** a_t in period t is the binary choice of whether to place an order:

$$a_t \in \mathcal{A} = \{0, 1\}.$$

If $a_t = 0$, we do not order. If $a_t = 1$, we place an order, which arrives immediately with probability 0.5 (we denote this arrival as $A_t = 1$) or fails to arrive with probability 0.5 ($A_t = 0$).

2 Finite-Horizon Dynamic Programming

2.1 Solution Method

To find the optimal policy, we solve the problem using dynamic programming with backward induction. We define $V_t(x)$ as the maximum expected total profit from the start of period t until the end of the horizon, given we start in state x.

The terminal condition is that at $t = 151$, the game is over and leftover inventory has no value:

$$V_{T+1}(x) = V_{151}(x) = 0, \quad \forall x \in \mathcal{X}$$

The Bellman equation defines the recursive relationship. The value $V_t(x)$ is the immediate, deterministic holding cost (paid at the start of the period) plus the maximum expected value of the two possible actions:

$$V_t(x) = \underbrace{-0.1 \cdot x}_{\text{Holding Cost}} + \max\left\{\mathbb{E}[\text{Profit } | a_t = 0], \mathbb{E}[\text{Profit } | a_t = 1]\right\}$$

Let $p_t = P(D_t = 1) = (t-1)/149$. The expected profit for each action is calculated by summing over all possible random outcomes.

2.1.1 Value of "Don't Order" ($a_t = 0$)

If we don't order, $A_t = 0$. The only randomness is demand D_t.

$$\begin{aligned}\mathbb{E}[\text{Profit } | a_t = 0] = \quad & (1 - p_t) \cdot \left[\underbrace{\min(x, 0)}_{\text{Revenue=0}} + \underbrace{V_{t+1}(x)}_{\text{Next State}=x}\right] \quad (\text{Case: } D_t = 0) \\ & + (p_t) \cdot \left[\underbrace{\min(x, 1)}_{\text{Revenue}} + \underbrace{V_{t+1}(\max(0, x-1))}_{\text{Next State}}\right] \quad (\text{Case: } D_t = 1)\end{aligned}$$

2.1.2 Value of "Order" ($a_t = 1$)

If we order, we face two independent random events: Arrival A_t and Demand D_t. This gives 4 outcomes. We can simplify this by noting that the "Order Fails" ($A_t = 0$) case is identical to the $a_t = 0$ case.

$$\mathbb{E}[\text{Profit } |a_t = 1] = 0.5 \cdot \mathbb{E}[\text{Profit } |a_t = 0] + 0.5 \cdot \mathbb{E}[\text{Profit } |a_t = 1, A_t = 1]$$

The "Order Arrives" ($A_t = 1$) term, where our inventory for the period is $x + 1$, is:

$$\begin{aligned} \mathbb{E}[\text{Profit } |a_t = 1, A_t = 1] = \quad & (1 - p_t) \cdot \left[\underbrace{\min(x + 1, 0)}_{\text{Revenue=0}} + \underbrace{V_{t+1}(x + 1)}_{\text{Next State}} \right] \quad \text{(Case: } D_t = 0\text{)} \\ & +(p_t) \cdot \left[\underbrace{\min(x + 1, 1)}_{\text{Revenue=1}} + \underbrace{V_{t+1}(x)}_{\text{Next State}} \right] \quad \text{(Case: } D_t = 1\text{)} \end{aligned}$$

We solve this system backward from $t = 150$ to $t = 1$ in a vectorized way to efficiently compute $V_t(x)$ and the optimal policy $\pi(t, x)$ for all states.

2.2 Results and Interpretation

After running the backward induction we get the value at our starting state is $\mathbf{V_1(5) = 32.368}$. The optimal policy $\pi(t, x)$ is shown in Figure 3.1 in the Appendix. It essentially says :"Given I am at time t (y-value) and have x items (x-value), what should I do?" 3 cases are detailed below.

- **Overall Structure:** The policy is a "threshold" policy. We order (black) if inventory is below a certain level, and do not order (white) if it is above that level. This makes sense, as ordering when we have high inventory would only incur holding costs.
- **Impact of Time (Seasonality):** The most important insight is how this threshold changes.
  1. **Early Periods ($t < 50$):** The policy is **cautious**. The demand probability p_t is very low, so a sale is rare. The profit from a sale (+1) does not justify the certain holding cost (-0.1). Therefore, the policy orders only when inventory is extremely low (e.g., $x \leq 2$).
  2. **Peak Season ($50 < t < 145$):** The policy becomes **aggressive**. As p_t (demand probability) increases, the expected profit from a sale becomes very high, easily outweighing the holding cost. The policy now orders even at high inventory levels ($x = 5$ or $x = 6$) because it is highly probable the item will be sold.
  3. **End of Horizon ($t \to 150$):** The policy becomes **cautious again**. The black "order" band shrinks. This is the "end-of-horizon" effect. Since $V_{151} = 0$ (leftovers are worthless), it becomes too risky to order. There is a 50% chance the order fails, and even if it arrives, you might pay a holding cost for an item you can't sell before the time runs out.

3 Simulation

3.1 Simulation Method

To validate our DP model, we conduct a forward simulation. We simulate 1000 independent "runs" of the 150-day season. Each simulation starts at $t = 1$ with $x_1 = 5$.

In each time step t, the "agent" looks at its current state x_t and consults the optimal policy $\pi(t, x_t)$ (calculated in Part B) to choose an action a_t. The environment then "rolls the dice" for demand D_t and (if $a_t = 1$) arrival A_t to determine the *actual* reward and next state x_{t+1}. We record the total cumulative reward for each of the 1000 runs.

3.2 Results and Interpretation

The distribution of the 1000 total rewards is shown in the histogram in Figure 3.2 in the Appendix.

- **Average Reward:** The mean reward over 1000 simulations was **32.352**.

Interpretation: The histogram shows the range of possible outcomes. Due to the randomness of demand and delivery, some runs resulted in a low total reward (≈ 25) and some in a high reward (≈ 40).

The most important result is that the **average reward from the simulation (32.352) is extremely close to the expected maximal reward calculated by our DP model (32.368)**. This validates that our DP model and policy are correct. The red dashed line in the histogram shows our calculated expected value, which sits perfectly at the center of the simulated distribution.

Appendix

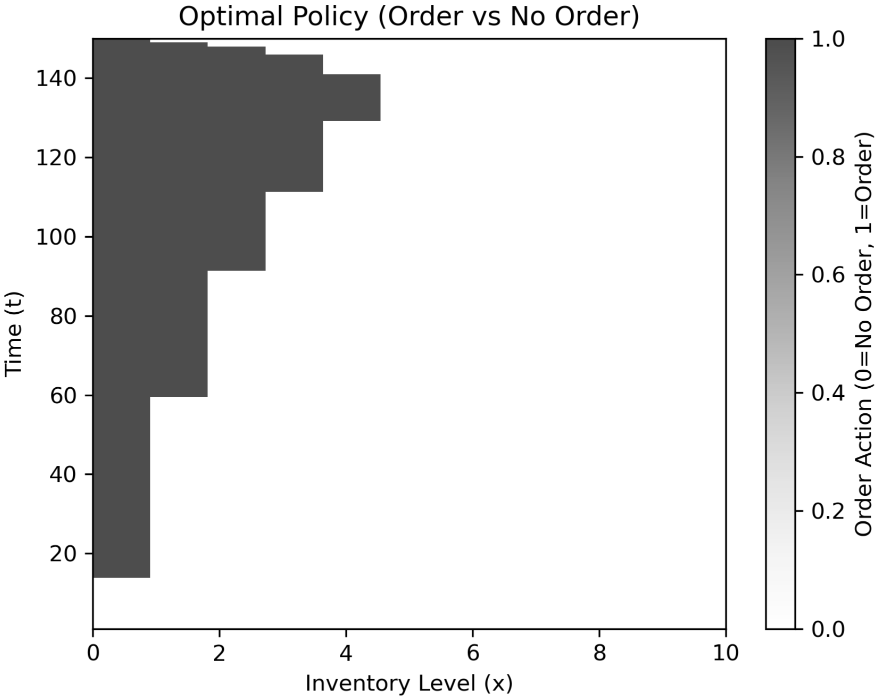

Figure 3.1: The optimal policy $\pi(t, x)$. The y-axis is the inventory level x_t and the x-axis is the time period t. Black (1) indicates the optimal action is "Order," and White (0) indicates "No Order."

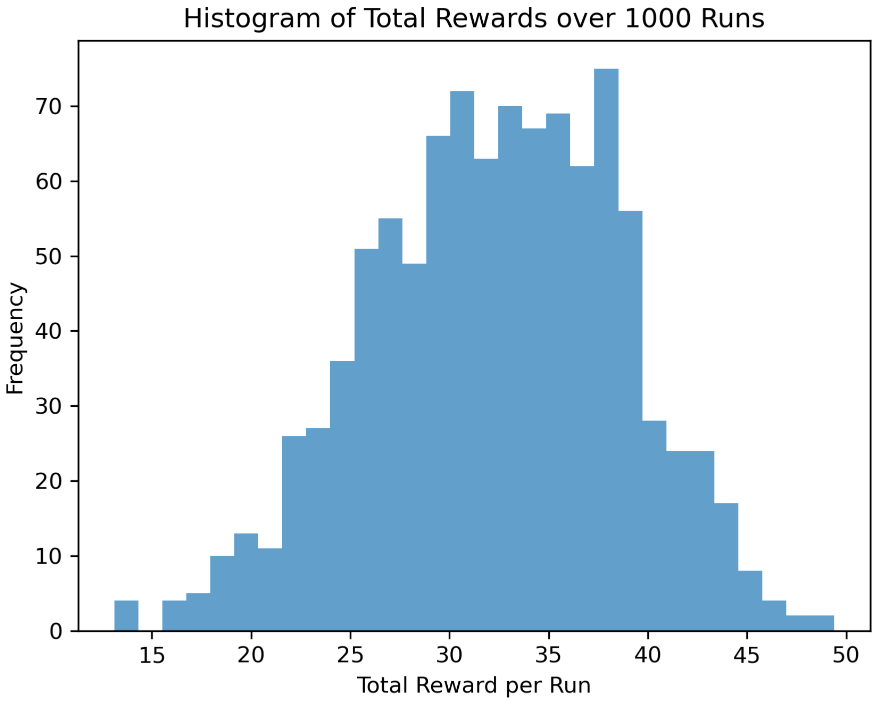

Figure 3.2: Histogram of total rewards from 1000 simulations following the optimal policy. The red dashed line indicates the theoretical expected reward $V_1(5)$ calculated by the DP algorithm.