Preliminaries

Initial Notation

- Simplex: Δ_n is the simplex of dimension n.
- States: $\mathcal{S} = \{s_1, \ldots, s_n\}$
- Actions: $\mathcal{A} = \{a_1, \ldots, a_m\}$
- State transitions:

  $G_{t,k}(i,j) = \text{prob}(X_{t+1} | X_t = s_i, U_t = a_k)$

  $G_{t,k} \geq 0, \quad G_{t,k}\mathbf{1} = \mathbf{1}$ (row stochastic)
- Markovian Policy
  - $\pi_t(s,a) \triangleq \text{prob}(U_t = a | X_t = s)$
  - $K_t(i,k) \triangleq \pi_t(s_i, a_k), \qquad K_t \geq 0, \quad K_t\mathbf{1} = \mathbf{1}$

    $\pi = (K_0, K_1, \ldots)$ (policy)
- Reward and Performance Metric
  - $R_t(s,a)$ (immediate reward)
  -

$$v^\pi = \mathbb{E}^\pi_{p_o}\left[\sum_{t=0}^{N-1} \gamma^t R_t(X_t, U_t) + \gamma^N r_N(x_N)\right] \tag{1}$$

$$= \mathbb{E}^\pi_{p_0}\left[\sum_{t=0}^{\infty} \gamma^t R_t(X_t, U_t)\right] \tag{2}$$

where $p_o \in \Delta_n$ is the probability density at $t = 0$ over states and $\gamma \in (0,1]$ is a discount factor. $\gamma < 1$ when $N = \infty$.

Observations

1. $p_t \in \Delta_n$ is a prob. dist. over states at time t.
   Given a decision policy, π

$$p_{t+1} = M^T_{\pi,t} p_t \tag{3}$$

where

$$M_{\pi,t} = \sum_{k=1}^{m} G_{t,k} \odot \left(K_t e_k \mathbf{1}^T\right) \qquad t = 0, 1, \ldots \tag{4}$$

where e_k is the k-th standard basis vector and $\odot$ is the Hadamard product that corresponds to element-wise multiplication. The above propogation defines a Markov Chain (MC) for the time evolution of the density. Indeed, $M_{\pi,t}$ is a Markov transition matrix whose element $[M_{\pi,t}]_{ij}$ is the probability of transitioning from state i to state j at time t under policy π. Note that $M_{\pi,t}$ is row stochastic.

$$M_{\pi,t} \geq 0, \quad M_{\pi,t}\mathbf{1} = \mathbf{1} \qquad\qquad p_{t+1}^T = p^T M_{\pi,t}. \tag{5}$$

2. Define the quantities

$$r_{\pi,t} \triangleq \big(R_t \odot K_t\big)\mathbf{1} \in \mathbb{R}^n \tag{6}$$

$$v^\pi = \sum_{t=0}^{N \text{ or } \infty} \gamma^t p_t^T r_{\pi,t} \tag{7}$$

   where p_t evolves according to to (5)

3. For $N < \infty$, finite, letting $V_{\pi,N} = r_N \in \mathbb{R}^n$.

$$V_{\pi,t} = r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1}, \qquad t = N-1, ..., 1, 0 \tag{8}$$

   We call $V_{\pi,t} \in \mathbf{R}^n$ the value function at time t. Note that that the value function is initialized as the expected reward at the final time and then propogates backwards in time according to the dynamics. The element $(V_{\pi,t})_i$ can be thought of as the "reward-to-go" from state i at time t when policy π is employed.

4. When a fixed set of actions are given, $\{U_0, U_1, \dots\}$, i.e. $U_t = a_k(t),\ t = 0, 1, \dots$ it can still be expressed as $\pi = (K_0, K_1, \dots)$ with $K_t = \mathbf{1}e_{k(t)}^T,\ t = 0, 1, \dots$

5. In all the following, we'll focus on Markovian policies, i.e. we do *not* consider history dependent ones. Policies can in general be:

   (a) Randomized history dependent (RHD)

   (b) Randomized Markovian (RM)

   (c) Deterministic history dependent (DHD)

   (d) Deterministic Markovian (DM)

   Note: Putterman's book has a proof of the fact that MDPs must have DM optimal policies when $\mathcal{S}$ and $\mathcal{A}$ are finite sets (see Proposition 4.4.3, p.90?) for MDPs with $N < \infty$. This is not necessarily true when the MDP has constraints!

Finite Horizon MDPs

Problem: $\max_\pi \ v^\pi$.

$$v^\pi = p_0^T r_{\pi,0} + \gamma p_1^T r_{\pi,1} + \cdots + \gamma^{N-1} p_{N-1}^T r_{\pi,N-1} + \gamma^N \underbrace{p_N^T r_N}_{p_{N-1}^T M_{\pi,N-1} V_{\pi,N}} \tag{9}$$

$$= \cdots + \gamma^{N-1} p_{N-1}^T \Big(\underbrace{r_{\pi,N-1} + \gamma M_{\pi,N-1} V_{\pi,N}}_{V_{\pi,N-1}} \Big) \tag{10}$$

$$\vdots \qquad \vdots \qquad \vdots \qquad \vdots$$

$$v^\pi = p_0^T V_{\pi,0} \tag{11}$$

Problem: $\max_\pi p_0^T V_{\pi,0}$
Since $V_{\pi,N} = r_N$ for all π, we can compute $V_{\pi,t}$ by backwards induction.

$$V_{\pi,t} = r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1} \qquad t = N-1, \ldots, 0 \tag{12}$$

Then note that $M_{\pi,t}$ and $r_{\pi,t}$ are linear in π_t and

$$e_j^T V_{\pi,t} = e_j^T r_{\pi,t} + \gamma e_j^T M_{\pi,t} \underbrace{V_{\pi,t+1}}_{:=y} \tag{13}$$

where

$$e_j^T r_{\pi,t} = e_j^T (R_t \odot K_t) \mathbf{1} \tag{14}$$

$$= (e_j^T R_t \odot \underbrace{e_j^T K_t}_{\substack{j\text{−th row} \\ \text{of } K_t}}) \mathbf{1} \tag{15}$$

$$e_j^T M_{\pi,t} y = e_j^T \left(\sum_{k=1}^m G_{t,k} \odot \left(K_t e_k \mathbf{1}^T \right) \right) y \tag{16}$$

$$= \left(\sum_{k=1}^m e_j^T G_{t,k} \odot \Big(\underbrace{e_j^T K_t}_{\substack{j\text{−th row} \\ \text{of } K_t}} e_k \mathbf{1}^T \Big) \right) y \tag{17}$$

It follows that $e_j^T V_{\pi,t}$ linearly depends only on the j-th row of K_t (decision variable for policy). Hence in the absence of all other constraints we can solve the following problem to obtain an optimal policy:
For each $j = 1, \ldots, n$.

$$\max_{e_j^T K_t} \quad e_j^T V_{\pi,t} \qquad t = 0, \ldots, N-1 \tag{18}$$

$$\text{s.t.} \quad e_j^T K_t \geq 0, \quad e_j^T K_t \mathbf{1} = 1 \tag{19}$$

for each row of $K_t,\ t = 0, \ldots, N-1$.
Since the cost is linear in the j-th row one of the optimal choices for $e_j^T K_t$ is a vertex of the probability simplex in $\mathbb{R}^m$!

Note that the overall cost to be maximized is

$$\sum_{t=0}^{N} \gamma^t p_t^T r_{\pi,t} = p_o^T V_{\pi,0} \tag{20}$$

$$= \sum_{t=0}^{T-1} \gamma^t p_t^T r_{\pi,t} + p_T^T V_{\pi,T} \tag{21}$$

for any $T = 0, ..., N-1$. It follows that

$$\max_{\pi} \quad \sum_{t=0}^{N} \gamma^t p_t^T r_{\pi,t} = \max_{\substack{K_0,\ldots,K_{T-1} \\ K_T,\ldots,K_{N-1}}} \sum_{t=0}^{T-1} \gamma^t p_t^T r_{\pi,t} + p_T^T V_{\pi,T} \tag{22}$$

Note p_T is independent of $K_T, \ldots, K_{N-1}$ and purely depends on $K_0, \ldots, K_{T-1}$, however $V_{\pi,T}$ is independent of $K_0, \ldots, K_{T-1}$.

In this case, meaningful problems are

$$\max_{\pi} \quad \sum_{t=0}^{N} \gamma^t p_t^T r_{\pi,t} \quad = \quad \max_{\pi} \quad p_0^T V_{\pi,0} \qquad\qquad \text{for a given } p_0. \tag{23}$$

OR

$$\max_{\pi} \; \min_{p_0} \quad \sum_{t=0}^{N} \gamma^t p_t^T r_{\pi,t} \quad = \quad \max_{\pi} \; \min_{p_0} \quad p_0^T V_{\pi,0} \tag{24}$$

Since $e_j^T V_{\pi,t}$ linearly depends on $e_j^T K_t$ only, it can be maximized with it's choice. Also, since

$$e_j^T V_{\pi,t} = e_j^T \underbrace{r_{\pi,t}}_{\substack{\text{linear} \\ \text{in } e_j^T K_t}} + \gamma e_j^T \underbrace{M_{\pi,t}}_{\substack{\text{linear} \\ \text{in } e_j^T K_t}} V_{\pi,t+1} \tag{25}$$

No matter how K_t is chosen to maximize $e_j^T V_{\pi,t}$, we have to maximize each component of $V_{\pi,t+1}$ separately simply since it's possible as above. More precisely,

$$\pi^* = \arg \max_{\pi} \min_{p_0 \in \Delta_n} \quad p_0^T V_{\pi,0} \tag{26}$$

$$= \arg \max_{\pi} \quad e_j^T V_{\pi,0} \qquad\qquad j = 1, \ldots, n \tag{27}$$

since $e_j \in \Delta_n$ for $j = 1, \ldots, n$. It follows that

$$V_{\pi^*,0} \geq V_{\pi,0} \qquad\qquad \forall \pi \tag{28}$$

$$\Rightarrow \qquad p_0^T V_{\pi^*,0} \geq p_0^T V_{\pi,0} \qquad\qquad \forall \pi \tag{29}$$

$$\Rightarrow \qquad \pi^* = \max_{\pi} \quad p_0^T V_{\pi,0} \qquad\qquad \forall p_0 \in \Delta_n \tag{30}$$

It follows that

$$\max_{\pi} \min_{p_0 \in \Delta} \quad p_0^T V_{\pi,0} \tag{31}$$

is a proper problem!
Now, recall that

$$V_{\pi,N} = r_N \tag{32}$$
$$V_{\pi,t} = r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1} \tag{33}$$

Since, as shown above, we have that

$$V_{\pi,t} = \begin{bmatrix} \max_\pi \; e_1^T V_{\pi,t} \\ \vdots \\ \max_\pi \; e_n^T V_{\pi,t} \end{bmatrix} \tag{34}$$

for a given $V_{\pi,t+1}$ we have that

$$V_{\pi^*,t} \geq V_{\pi,t} \qquad \forall \pi, \quad t = 0, \dots, N-1 \tag{35}$$

The conclusion is that no matter how p_t evolves in time, the optimal policy can be computed in one-shot via dynamic programming (DP).

Dynamic Programming for Finite-Horizon MDPs

- Initialize. $V_N^* = r_N$
- For $t = N-1, \dots, 0$, iteratively compute

$$e_j^T K_t^* = \quad \arg \max_{e_j^T K_t^*, e_j \in \Delta_n} \quad e_j^T r_{\pi,t} + e_j^T \gamma M_{\pi,t} V_{\pi,t+1}^* \tag{36}$$
$$V_{\pi,t}^* = \quad \max_{e_j^T K_t^*, e_j \in \Delta_n} \quad e_j^T r_{\pi,t} + e_j^T \gamma M_{\pi,t} V_{\pi,t+1}^* \tag{37}$$

Since the above maximization is an LP over Δ_n, we can always find an optimal policy that is deterministic Markovian (DMP) in the absence of constraints.

Infinite Horizon MDPs

In the infinite horizon case, we assume the transition kernel and rewards are constant over time, i.e. $G_t = G$ and $R_t = R$ for all t.

Theorem 1 (Existence and Uniqueness)
The following equation has a unique solution

$$T(V) = \max_{\pi} \Big(r_\pi + \gamma M_\pi V \Big) \tag{38}$$

Proof 1 *The proof proceeds by using the Banach fixed point theorem on the Bellman operator $T(V)$. Consider $V_1 \to V_1^*$ and $V_2 \to V_2^*$ for any $V_1, V_2 \in \mathbb{R}^n$*

$$V_k^* = \max_\pi (r_\pi + \gamma M_\pi V_k), \qquad k = 1, 2 \tag{39}$$

We consider the difference $||T(V_1) - T(V_2)||_\infty$ under the infinity norm.

$$\begin{aligned}
||T(V_2) - T(V_1)||_\infty &= \left|\left|\max_\pi (r_\pi + \gamma M_\pi V_2) - \max_\pi (r_\pi + \gamma M_\pi V_1)\right|\right|_\infty && (40)\\
&= \max_{j=1\dots n} \left|\max_\pi e_j^T (r_\pi + \gamma M_\pi V_2) - \max_\pi e_j^T (r_\pi + \gamma M_\pi V_1)\right| && (41)\\
&\le \max_j \max_\pi \left|e_j^T (r_\pi + \gamma M_\pi V_2) - e_j^T (r_\pi + \gamma M_\pi V_1)\right| && (42)\\
&\le \max_j \max_\pi \gamma \left|e_j^T M_\pi V_2 - e_j^T M_\pi V_1\right| && (43)\\
&= \max_\pi \gamma \, ||M_\pi (V_2 - V_1)||_\infty && (44)\\
&= \max_\pi \gamma \, ||V_2 - V_1||_\infty && (45)
\end{aligned}$$

where we have used two facts:

$$\max_x f(x) - max_y g(y) \le \max_x f(x) - g(x) \tag{46}$$

and

$$||Mv||_\infty \le ||M||_\infty \, ||v||_\infty = ||v||_\infty \tag{47}$$

since $||M||_\infty = \max_j ||e_j^T M||_1 = 1$. Thus we have that $T(V)$ is a contractive mapping for $\gamma \in [0, 1)$. By the Banach fixed point theorem, it follows that $V = T(V)$ has a unique solution V^* and that the sequence $V_{k+1} = T(V_k)$ for any V_0 will converge to V^*.

Explicitly we have that

$$\begin{aligned}
&& ||V_{k+1} - V_k|| \le \gamma ||V_k - V_{k-1}|| && (48)\\
\Rightarrow && ||V_k - V^*|| \le \gamma^k ||V_0 - V^*|| && (49)
\end{aligned}$$

This proof leads to a straight-forward technique for computing the value function called *value iteration.*

Value Iteration

- Pick V_0.
- Compute $V_{k+1} = T(V_k), \; k = 0, 1, \dots$.
- Stop when $||V_{k+1} - V_k||$ is within some desired tolerance.

Properties of V^* and the corresponding π^*:

$$v^\pi = \textstyle\sum_{t=0}^\infty \gamma^t \underbrace{p_t^T}_{\text{linear in } \pi} \overbrace{r_{\pi,t}}^{\text{linear in } \pi} \tag{50}$$

For any policy, since $\exists \alpha_1, \alpha_2 > 0$ such that

$$||r_{\pi,t}|| \leq \alpha_1, \qquad ||p_t|| \leq \alpha_2 \qquad \Rightarrow \qquad |p_t^T r_{\pi,t}| \leq \underbrace{\alpha_1\alpha_2}_{:=\alpha} \tag{51}$$

$$\Rightarrow \qquad \sum_{t=0}^\infty \gamma^t |p_t^T r_{\pi,t}| = \alpha \sum_{t=0}^\infty \gamma^t = \alpha \frac{1}{1-\gamma} \tag{52}$$

It follows that v^π is absolutely convergent and thus $v^\pi = c$ for some c, i.e. all policies have a finite reward for any $p_0 \in \Delta_n$.

$$p_t^T = p_0^T M_{\pi,0} \cdots M_{\pi,t-1} \tag{53}$$

$$v^\pi = \sum_{t=0}^\infty \gamma^t p_0^T \Big(\underbrace{M_{\pi,0} \cdots M_{\pi,t-1}}_{:=\tilde{M}_{\pi,t-1}} \Big) r_{\pi,t} \tag{54}$$

$$= p_0^T \Big[r_{\pi,0} + \gamma M_{\pi,0} r_{\pi,1} + \gamma^2 M_{\pi,0} M_{\pi,1} r_{\pi,2} + \cdots \Big] \tag{55}$$

$$= p_0^T \Big[r_{\pi,0} + \gamma M_{\pi,0} \Big(\underbrace{r_{\pi,1} + \gamma M_{\pi,1} r_{\pi,2} + \cdots}_{:=V_{\pi,1}} \Big) \Big] \tag{56}$$

$$= p_0^T \Big[\underbrace{r_{\pi,0} + \gamma M_{\pi,0} V_{\pi,1}}_{:=V_{\pi,0}} \Big] \tag{57}$$

where

$$V_{\pi,1} = r_{\pi,1} + \gamma M_{\pi,1} \Big(\underbrace{r_{\pi,l2} + \gamma M_{\pi,2} r_{\pi,3} + \cdots}_{:=V_{\pi,2}} \Big) \tag{58}$$

$$\Rightarrow \qquad V_{\pi,t} \triangleq r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1} \tag{59}$$

Note $V_{\pi,t}$ are well-defined for all $\gamma \in [0,1)$ due to absolute convergence. It follows that $v^\pi = p_0^T V_{\pi,0}$. As before $\max_\pi v^\pi = \max_\pi e_j^T V_{\pi,0}$ $j = 1, \ldots, n$.

$$V_{\pi,0} = r_{\pi,0} + \gamma M_{\pi,0} V_{\pi,1} \tag{60}$$

$$\vdots \tag{61}$$

$$V_{\pi,t} = r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1} \qquad t = 0, 1, 2, \ldots \tag{62}$$

Note: Consider a stationary policy, $\pi = (\pi, \pi, \dots)$ with abuse of notation

$$V_\pi = r_\pi + \gamma M_\pi V_\pi \tag{63}$$

$$\Rightarrow V_\pi = (I - \gamma M_\pi)^{-1} r_\pi \tag{64}$$

The spectrum of M_π is in the unit circle $\Rightarrow I - \gamma M_\pi$ cannot have a zero eigenvalue. Thus $I - \gamma M_\pi$ is invertible and for stationary policies for stationary processes, we have that

$$V_\pi = (I - \gamma M_\pi)^{-1} r_\pi \tag{65}$$

$$= \sum_{t=0}^{\infty} \gamma^t M_\pi^t r_\pi \tag{66}$$

Since, for Markovian policies,

$$V_{\pi,t} = r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1} \tag{67}$$

For optimal, Markovian policies

$$V_{\pi,t}^* = \max_\pi r_{\pi,t} + \gamma M_{\pi,t} V_{\pi,t+1}^* \tag{68}$$

i.e. $V_{\pi,t}^* = T(V_{\pi,t+1}^*)$.

Finite Horizon Total Reward Linear Programming

Finite horizon, $\gamma = 1$.

Primal Problem

$$\min_{\{W_t\}_{t=0}^N} \quad p_0^T W_0 \tag{69}$$

$$\text{s.t.} \quad W_N \geq R_n e_k, \tag{70}$$

$$W_t[i] \geq \sum_j G_{t,k}[i,j]\Big(W_{t+1}[j] + R_t(i,k)\Big) \qquad \forall i,\ \forall a_k,\ t = N-1, \dots, 0 \tag{71}$$

Matrix form:

$$\min_{\{W_t\}_{t=0}^N} \quad p_0^T W_0 \tag{72}$$

$$\text{s.t.} \quad W_N \geq R_n e_k, \tag{73}$$

$$W_t \geq G_{t,k} W_{t+1} + \Big(G_{t,k} \odot (R_t e_k \mathbf{1}^T)\Big)\mathbf{1} \qquad \forall a_k,\ t = N-1, \dots, 0 \tag{74}$$

Variables:

- $W_t \in \mathbb{R}^n$: Upper bound on the reward-to-go.
- W_t^* is the reward-to-go
- p_0: initial distribution, must have positive mass in all states.

Dual Problem

$$\begin{aligned} \max_{\{y_t\}_{t=0}^N} \quad & \sum_t \mathbf{1}^T \big(R_t \odot y_t\big)\mathbf{1} && (75) \\ \text{s.t.} \quad & y_0 = p_0, \quad y_t \geq 0 \qquad\qquad t = 0, \ldots, N && (76) \\ & \sum_k y_{t+1}[i,k] = \sum_i \sum_k G_{t,k}^T[i,j] y_t[j,k] && (77) \end{aligned}$$

$$\begin{aligned} \max_{\{y_t\}_{t=0}^N} \quad & \sum_t \mathbf{1}^T \big(R_t \odot y_t\big)\mathbf{1} && (78) \\ \text{s.t.} \quad & y_0 = p_0, \quad y_t \geq 0 \qquad\qquad t = 0, \ldots, N && (79) \\ & y_{t+1}\mathbf{1} = \sum_k \Big(G_{t,k}^T \odot \mathbf{1} e_k^T y_t^T\Big)\mathbf{1} && (80) \end{aligned}$$

Congestion Game

- population of infinitesimal agents
- $z_t[s_i, a_k]$: pop. in state s_i choosing a_k at time t
- Competition:
  - $R_t[s_i, a_k](z_t)$: rewards depend on population
  - usually $R_t[s_i, a_k]\Big(z_t(s_i, a_k)\Big)$.
  - usually $R_t[s_i, a_k](u)$ is a decreasing function of mass, i.e. "congestion".

Recall that for a policy π, $(K_0, \ldots, K_N)$...

$$r_{\pi,t} = \Big(R_t(z_t) \odot K_t\Big)\mathbf{1}, \qquad v^\pi = \sum_{t=0}^N \gamma^t p_t^T r_{\pi,t} = \sum_{t=0}^N p_t^T r_{\pi,t} \tag{81}$$

Equilibrium Concept (Wardrop Equilibrium)

Wardrop Equilibrium = "continuous population Nash"

Definition 1 *A population distribution $z_0, \ldots, z_N$ is a* Wardrop equilibrium *if*

$$v^\pi \geq v^{\pi'} \tag{82}$$

for all π sucht that $\pi_t(i,k) > 0$ only if $z_t(i,k) > 0$.

Intuitively, this means any strategy that is positively support by z_t at each time step is optimal, i.e. no population mass chooses an inefficient strategy.

Potential Game

Definition 2 *$\exists$ a function $F(z) \in \mathbf{C}^1$ such that*

$$\frac{\partial F}{\partial z_t(i,k)} = R_t[i,k](z_t) \tag{83}$$

In the standard case, $R_t[i,k](z_t) = R_t[i,k]\Big(z_t(s,a)\Big)$ and we can write a potential function similar to the Rosenthal potential from non-atomic routing games

$$F(z) = \sum_t \sum_i \sum_k \int_0^{z_t(i,k)} R_t[i,k](u)\ du \tag{84}$$

Potential Optimization Problem

Initial population distribution: $m_0 \in \mathbb{R}^n$

$$\begin{aligned} \max_{\{z_t\}_{t=0}^N} \quad & F(z) \\ \text{s.t.} \quad & z_0 = m_0, \quad z_t \geq 0 \qquad\qquad t = 0, \ldots, N \\ & \sum_k z_{t+1}[i,k] = \sum_i \sum_k G_{t,k}^T[i,j] z_t[j,k] \end{aligned} \tag{85}$$

Matrix form:

$$\begin{aligned} \max_{\{z_t\}_{t=0}^N} \quad & F(z) \\ \text{s.t.} \quad & z_0 = m_0, \quad z_t \geq 0 \qquad\qquad t = 0, \ldots, N \\ & z_{t+1}\mathbf{1} = \sum_k \Big(G_{t,k}^T \odot \mathbf{1} e_k^T z_t^T\Big)\mathbf{1} \end{aligned} \tag{86}$$

- $W_t \in \mathbb{R}^n$: Dual variables for mass propagation constraints.
- $\mu_t \in \mathbb{R}^{n\times m}$: Dual variables for positivity constraints.

Theorem 2 *A minimizer of* (85) *is a* Wardrop Equilibrium*.*

Proof 2 *Equilibrium conditions:*

$$\underbrace{R_t[i,k]}_{\frac{\partial F}{\partial z_t[i,k]}} - W_t[i] + e_i^T G_{t,k} W_{t+1} + \mu_t[i,k] = 0 \tag{87}$$

$$R_N[i,k] - W_N[i] + \mu_N[i,k] = 0 \tag{88}$$

$$z_t \geq 0, \quad \mu_t \geq 0, \quad z_t[i,k]\mu_t[i,k] = 0 \tag{89}$$

For a given π, $(K_0, \dots, K_N)$, we want to compute the cost v^π using the optimality conditions as certificates on the cost.

$$
\begin{aligned}
v^\pi &= p_0^T r_{\pi,0} + \cdots + p_{N-1}^T r_{\pi,N-1} + \underbrace{p_N^T r_N}_{p_{N-1}^T M_{\pi,N-1} V_{\pi,N}} && (90)\\
&= \cdots + p_{N-1}^T \Big(\underbrace{r_{\pi,N-1} + M_{\pi,N-1} V_{\pi,N}}_{V_{\pi,N-1}} \Big) && (91)\\
&\vdots \qquad \vdots \qquad \vdots \qquad \vdots\\
v^\pi &= p_0^T V_{\pi,0} && (92)
\end{aligned}
$$

$$
\begin{aligned}
V_{\pi,N}[i] = r_{\pi,N}[i] &= \sum_k R_N[i,k] K_N[i,k] && (93)\\
&= \sum_k W_N[i] K_N[i,k] - \sum_k \mu_N[i,k] K_N[i,k] && (94)\\
&= W_N[i] - \underbrace{\sum_k \mu_N[i,k] K_N[i,k]}_{:=\tilde{\mu}_{\pi,N}[i]} && (95)
\end{aligned}
$$

$$
\begin{aligned}
V_{\pi,N-1}[i] &= r_{\pi,N-1}[i] + e_i^T M_{\pi,N-1} V_{\pi,N} && (96)\\
&= \sum_k R_{N-1}[i,k] K_{N-1}[i,k] + \sum_j \sum_k G_{N-1,k}[i,j] K_{N-1}[i,k] V_{\pi,N}[j] && (97)\\
&= \sum_k K_{N-1}[i,k] \left(\underbrace{R_{N-1}[i,k]}_{\substack{W_{N-1}[i] \\ -e_i^T G_{N-1,k} W_N \\ -\mu_{N-1}[i,k]}} + e_i^T G_{N-1,k} \underbrace{V_{\pi,N}}_{\substack{W_N[i] \\ -\tilde{\mu}_{\pi,N}[i]}} \right) && (98)\\
&= \sum_k K_{N-1}[i,k] \Big(W_{N-1}[i] - \mu_{N-1}[i,k] - e_i^T G_{N-1,k} \tilde{\mu}_{\pi,N}[i] \Big) && (99)\\
&= W_{N-1}[i] - \underbrace{\sum_k K_{N-1}[i,k] \Big(\mu_{N-1}[i,k] + e_i^T G_{N-1,k} \tilde{\mu}_{\pi,N}[i] \Big)}_{:=\tilde{\mu}_{\pi,N-1}[i]} && \vdots\\
& && (100)\\
V_{\pi,0} &= W_0 - \tilde{\mu}_{\pi,0} && (101)
\end{aligned}
$$

Note that $\tilde{\mu}_{\pi,t} \geq 0$. If $K_t[i,k] > 0$ only when $z_t[i,k] > 0$, it follows by complementary slackness that $K_t[i,k] > 0$ only when $\mu_t[i,k] = 0$. It follows that $\tilde{\mu}_{\pi,t} = 0$ for all t.

Thus for π that satisfies the equilibrium condition,

$$v^{\pi} = p_0^T V_{\pi,0} = p_0^T W_0 \geq p_0^T W_0 - p_0^T \tilde{\mu}_{\pi',0} = v^{\pi'} \tag{102}$$