Markov Decision Process Visualizations

Dan Calderone

***Abstract*—This tutorial details matrix notation and visualizations for stochastic network flow and optimization problems, ie. Markov decision processes (MDPs). We review an incidence matrix style notation and as well as affine and vertex constraint representations for Markov chains and Markov decision processes in the average-reward, discounted-reward, and total-reward versions of the problem in both the infinite and finite horizon cases. We review primal and dual linear programming formulations of each of these problems along with variable interpretations. Detailed visualizations of the geometry of these problems are provided throughout.**

I. Introduction

Algebraic graph theory has become a staple modern engineering problems. In this tutorial paper, we review basic algebraic grapth theory constructions with references and visualizations. We then present mass flow constraint formulations standard to many modern optimization problems as well as linear programming formulations associated with shortest path problems. We note that while the visualizations in this paper are more extensive and thorough than normal, virtually none of the mathematical content is original; we have sought to provide the appropriate references throughout.

Markov decision process have become a staple of modern machine learning problems. In this tutorial paper, we collect a matrix notation for Markov decision processes, present the well known MDP linear programming formulations in terms of this notation, and provide a thorough set of visualizations of both the primal and dual linear programs. The matrix notation also allows us to make connections with the underlying graph structure of the MDP and several algebraic graph theory constructions (most notably the node-edge incidence matrix of a directed graph).

Markov decision processes (MDPs) are a staple of modern machine learning and control theory used for modeling discrete decision processes with a discrete state space. In this paper, we review and consolidate the excellent matrix notation for MDPs presented in [1] and make connections with standard algebraic graph theory constructions (most notably the node-edge incidence matrix of a directed graph.) We present well-known linear programming formulations of MDPs in the infinite horizon (average-reward and discounted reward) and finite horizon (total reward) settings [2] in both their primal and dual forms. Exposition, compact proofs, and minor extensions are offered throughout. Extensive visualizations are included throughout.[1]

The paper is organized as follows. In Section **??**, we define notational preliminaries and present several illustrative examples of our visualization techniques. (For more thorough explanation of the visualization techniques, we suggest the linear programming tutorial offered by these authors as well [**?**].)

In Section II, we present notation and visualizations for stochastic transition kernels, policies, and the resulting Markov chains. We also relate these concepts algebraically to the underlying graph structure. In Section **??**, we present the infinite horizon, average reward LP formulation of an MDP and provide visualizations. In Section **??**, we examine the effect of a discount factor on the transition kernel, present the infinite horizon, discounted reward LP formulation of an MDP and provide visualizations. In Section **??**, we present the finite horizon, total (and discounted) reward LP formulations of an MDP and provide visualizations. Appendix **??** contains a table summarizing notation for reference.

This paper assumes the knowledge and notation in the following monographs.

- Vector visualizations
- Matrix column geometry
- Linear programming geometry
- Graph and network optimization geometry

II. Stochastic Network Flows

Transition Kernel: A Markov decision processes consists of a state space $\mathcal{S}$, an action set $\mathcal{A}$, and a transition kernel $P \in [0,1]_{|\mathcal{S}|\times|\mathcal{A}|}$ that gives the probability of transitioning to a new state s' from state s when action a is taken. We will assume there is a unique set of actions associated with each state s, $\mathcal{A}_s$ and the full set of actions $\mathcal{A} = \cup_s \mathcal{A}_s$. We will use $a \in \mathcal{A}$ to index into state-action pairs (with the appropriate state s for a given action a implied.) Let $y \in \Delta_{|\mathcal{A}|}$ represent a mass distributions over the state-action pairs. We will use y_a to refer to the mass choosing action a (from state s).

The *Markov property* states that these transitions only depend on the current state. We will represent the transition kernel as a matrix $P \in [0,1]^{|\mathcal{S}|\times|\mathcal{A}|}$

$$\big[P\big]_{s',a} = \text{Prob}(s'|s,a)$$

For clarity we will assume that the first set of columns of P correspond to the actions from state 1, the second set of columns correspond to the actions from state 2, etc. We may also use $P_s \in [0,1]^{|\mathcal{S}|\times|\mathcal{A}_s|}$ to represent the sub matrix that gives the transition from state s to the other states for various actions within $\mathcal{A}_s$. With this notation, we have that

$$P = \begin{bmatrix} P_1 & \dots & P_{|\mathcal{S}|} \end{bmatrix}$$

We also define an indicator matrix $E_s \in \{0,1\}^{|\mathcal{S}|\times|\mathcal{A}|}$

$$\big[E_s\big]_{s',a} = \begin{cases} 1 & ; \text{ if } s = s' \\ 0 & ; \text{ otherwise} \end{cases}$$

The order of the columns of E_s should be consistent with the columns of P. Assuming the structure of P detailed above (and each action available from each state), $E_s = I_{|\mathcal{S}|\times|\mathcal{S}|} \otimes \mathbf{1}^T$.

Given the underlying graph structure we can define an edge/state-action incidence matrix $W \in [0,1]^{|\mathcal{E}|\times|\mathcal{S}||\mathcal{A}|}$ such that

$$\big[W\big]_{e,a} = \begin{cases} P_{s',a}; & \text{if } e \text{ runs from } s \text{ to } s' \\ 0 & ; \text{ otherwise} \end{cases} \tag{1}$$

We then have the following identities

$$E_o W = E_s, \qquad E_i W = P \tag{2}$$

Mass balance over a stochastic transition network is given by $E_s y = P y$. Note that E_s, P, W are also column stochastic

$$\mathbf{1}^T E_s = \mathbf{1}^T, \quad \mathbf{1}^T P = \mathbf{1}^T, \quad \mathbf{1}^T W = \mathbf{1}^T \tag{3}$$

and mass conservation is given by

$$\mathbf{1}^T \rho = \mathbf{1}^T E_s y = \mathbf{1}^T P y = \mathbf{1}^T y \tag{4}$$

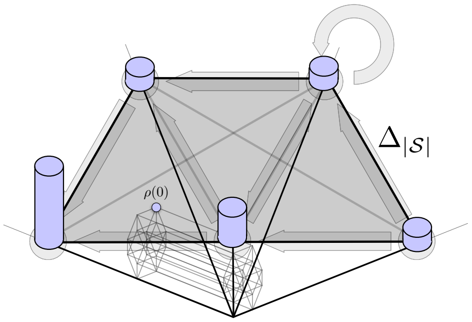

Fig. 1: State distribution visualization

Matrix: Transition Kernels

$$W \in \mathbb{R}^{|\mathcal{E}|\times|\mathcal{A}|} \qquad P = E_i W \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{A}|}$$

Ex:

$$\begin{array}{cc} \leftarrow \text{Actions} \rightarrow & \\ P = \begin{bmatrix} \\ P_1 & P_2 & P_3 & P_4 & P_5 \\ \\ \end{bmatrix} & \begin{matrix} \uparrow \\ \text{States} \\ \downarrow \end{matrix} \end{array}$$

with $P_2 = I_3; P_3 = I_4; P_5 = I_1$ and

$$P_1 = \begin{bmatrix} 0.13 & 0.32 & 0.32 \\ 0.70 & 0.63 & 0.63 \\ 0.17 & 0.05 & 0.05 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad P_4 = \begin{bmatrix} 0.13 & 0.32 & 0.32 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0.70 & 0.63 & 0.63 \\ 0.17 & 0.05 & 0.05 \end{bmatrix},$$

$$\begin{array}{cc} \leftarrow \text{Actions} \rightarrow & \\ W = \begin{bmatrix} 0.13 & 0.32 & 0.32 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.70 & 0.63 & 0.63 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.17 & 0.05 & 0.05 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.13 & 0.32 & 0.32 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.70 & 0.63 & 0.63 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0.17 & 0.05 & 0.05 & 0 \\ 0 & 0 & 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{bmatrix} & \begin{matrix} \uparrow \\ \text{Edges} \\ \downarrow \end{matrix} \end{array}$$

Properties: Column stochastic

Example

For visualization purposes, we consider the following transition kernel shown in the graph. P and W for this transition kernel are illustrated in the box.

Policies and Markov Chains

Flow over a stochastic network is determined by choosing a mixed strategy over actions at each state $\pi^s \in \Delta_{|\mathcal{A}_s|}$. Together these mixed strategies are referred to as a *policy*. We will use $\pi_a \in [0,1]$ to refer to the probability of choosing action a from state s and

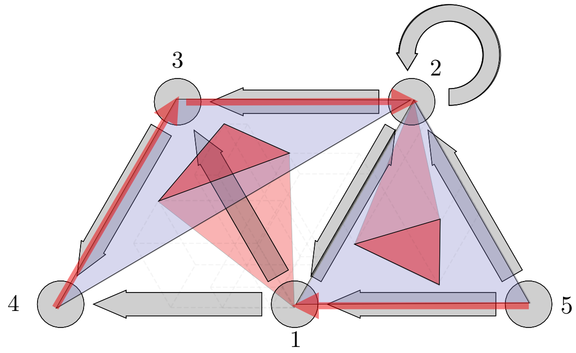

Fig. 2: Graph-policy structure

$\pi \in \times_s \Delta_{|\mathcal{A}_s|}$ to refer to the the full feedback policy. It is also convenient to organize the policy into a matrix $\Pi \in [0,1]^{|\mathcal{A}| \times |\mathcal{S}|}$

$$[\Pi]_{a,s} = \begin{cases} \pi_a & ; \text{ if } a \in \mathcal{A}_s \\ 0 & ; \text{ otherwise} \end{cases}$$

Matrix: Policy Matrix

$$\Pi \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{S}|}$$

Ex:

$$\Pi = \begin{array}{c} \leftarrow \text{States} \rightarrow \\ \begin{bmatrix} \pi_1 & 0 & \dots & 0 \\ 0 & \pi_2 & & \vdots \\ \vdots & & \ddots & \vdots \\ 0 & \dots & \dots & \pi_{|\mathcal{S}|} \end{bmatrix} \end{array} \begin{array}{c} \uparrow \\ \text{Actions} \\ \downarrow \end{array} = \begin{bmatrix} 0.1 & 0 & 0 & 0 & 0 \\ 0.2 & 0 & 0 & 0 & 0 \\ 0.7 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0.3 & 0 \\ 0 & 0 & 0 & 0.4 & 0 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}$$

Properties: Column stochastic

For clarity, given the above structure of P and E_s, Π has a block diagonal structure

Given π in vector form, note that $\Pi = \mathbf{d}(\pi) E_s^T$. Note also that Π is column stochastic.

$$\mathbf{1}^T \Pi = \mathbf{1}^T$$

A transition kernel can be thought of as sets of possible columns for a Markov transition matrix. A policy selects the actual columns of M from the convex hull of the possible columns given in the transition kernel, collapsing the transition kernel down to a single Markov chain. Indeed,a policy matrix Π satisfies the identities

$$E_s \Pi = I, \qquad P\Pi = M.$$

where $M \in [0,1]^{|\mathcal{S}| \times |\mathcal{S}|}$ is the Markov state-to-state transition matrix. Note with notation of Equation (1) we have that $M[:, s] = P_s \pi_s$. We can also define a Markov transition matrix from state-action to state-action pairs defined by a policy

$$N = \Pi P$$

Note that N is square, $N \in [0,1]^{|\mathcal{A}| \times |\mathcal{A}|}$, and also that in general, N is not full rank.

We say that π is a *pure strategy policy* if for each $s \in \mathcal{S}$, π^s puts all mass on a single action. We can denote the set of pure strategy policies as Γ. Note that $|\Gamma| = \prod_s |\mathcal{A}_s|$. For the sample transition kernel listed above there are six pure strategy policies.

$$\Pi_1 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_2 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_3 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

$$\Pi_4 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_5 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \\ 0 & 0 & 0 \end{bmatrix}, \quad \Pi_6 = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & 1 \end{bmatrix}$$

We make the following standard assumption for pure strategy policies.

Assumption 1. *For every pure strategy policy* Π*, the resulting Markov chain* $M = P\Pi$ *is aperiodic and irreducible.*

Note that this assumption is enough to guarantee the existence of a unique stationary distribution $\rho \in \Delta_{|\mathcal{S}|}$ for any policy π in matrix form Π.

$$\rho = M\rho = P\Pi\rho \tag{5}$$

Given a state distribution ρ, we can compute the joint-distribution over the state-action set $y \in \Delta_{|\mathcal{A}|}$ as

$$y = \Pi\rho$$

Note that the joint distribution can be expressed in terms of the state distribution as

$$\begin{aligned} \rho &= P\Pi\rho \\ E_s\Pi\rho &= P\Pi\rho \\ E_s y &= Py \end{aligned}$$

Note that y is unique given π and ρ. Given y, ρ is uniquely determined as $\rho = E_s y$. π is almost uniquely determined by

$$\pi_a = \frac{y_a}{\sum_{a \in \mathcal{A}_s} y_a} = \frac{y_a}{\rho_s}$$

Equation (6) becomes undetermined when $\rho_s = 0$; however, this is unimportant since the policy from a given state is only relevant when there is positive probability mass in that state. Assuming that $\rho > 0$, we have that

$$\Pi = \mathbf{d}(y) E_s \mathbf{d}\big(E_s y\big)^{-1} = \mathbf{d}(y) E_s \mathbf{d}\big(\rho\big)^{-1}$$

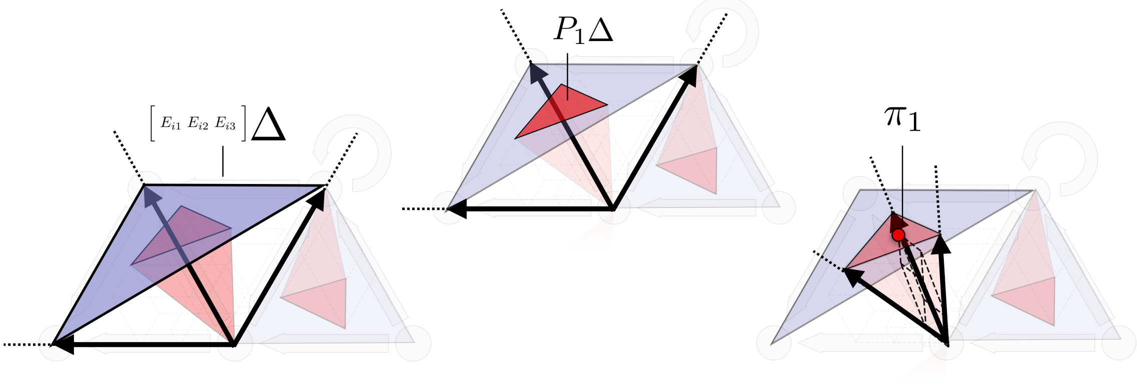
$\begin{bmatrix} E_{i1} & E_{i2} & E_{i3} \end{bmatrix} \Delta$
$P_1\Delta$
π_1

Fig. 3: Graph-policy structure

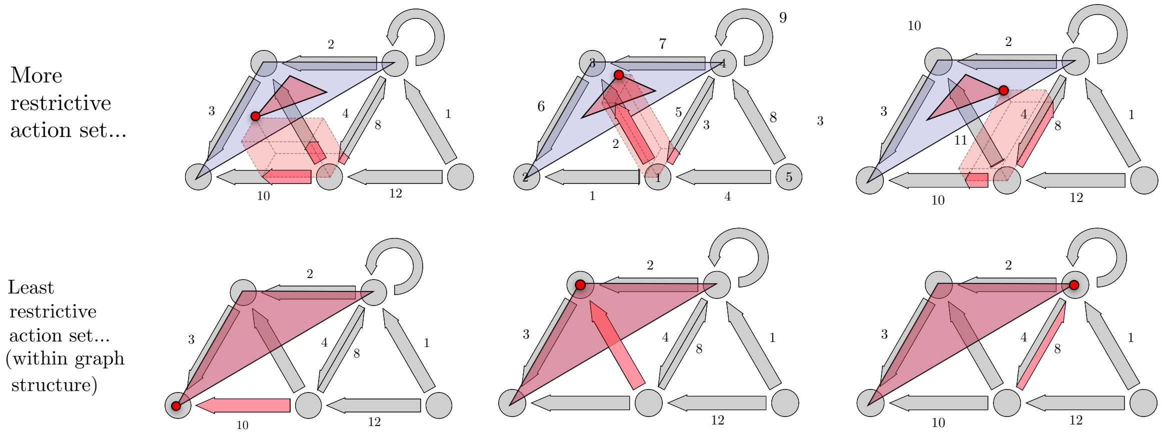
More restrictive action set...
Least restrictive action set... (within graph structure)
2
3
4
8
1
10
12
9
7
3
4
6
5
3
8
3
2
2
1
5
1
4
10
11

Fig. 4: Graph-policy example

A. *Markov Chain Time Evolution*

Markov transition matrices define the update equations

$$\rho(t{+}1) = M\rho(t), \quad \rho(0) \in \Delta_{|\mathcal{S}|} \tag{6}$$

$$y(t{+}1) = Ny(t), \quad y(0) \in \Delta_{|\mathcal{A}|} \tag{7}$$

for initial distributions $\rho(0)$ or $y(0)$. Since the initial state distributions lives in the probability simplex, $\rho(1)$ is contained in the convex hull of the columns of M, $\rho(2)$ is contained in the convex hull of the columns of M^2, etc. Similarly, $y(1)$ is contained in the convex hull of the columns of N, $y(2)$ is contained in the convex hull of the columns of N^2, etc. If the columns of M and N live on the interior of the simplex (a sufficient condition for Assumption 1), they represent contraction maps on the simplex and the corresponding fixed points (from Brouwer's fixed point theorem) are the steady state distributions. to Brouwer's fixed point theorem. One can visualizes the time evolution of the Markov chain as the simplex contracting down to the steady state distribution. This contraction process is illustrated in Fig. **??**.

In the joint distribution space, selecting a policy chooses a slice of the joint distribution space determined by the columns of Π. The columns of N are then given by the columns of P with the columns of Π treated as coordinate vectors. No matter what the initial joint distribution is, $y(t)$ for $t \geq 1$ will live in the convex hull of the columns of Π.

The convergence of the columns of M^t and N^t to the steady state distribution for each pure strategy policy are illustrated in Fig. **??** and Fig. **??**.

Note that a steady state distribution $\rho \in \Delta_{|\mathcal{S}|}$ satisfies

$$\rho = P\Pi\rho$$

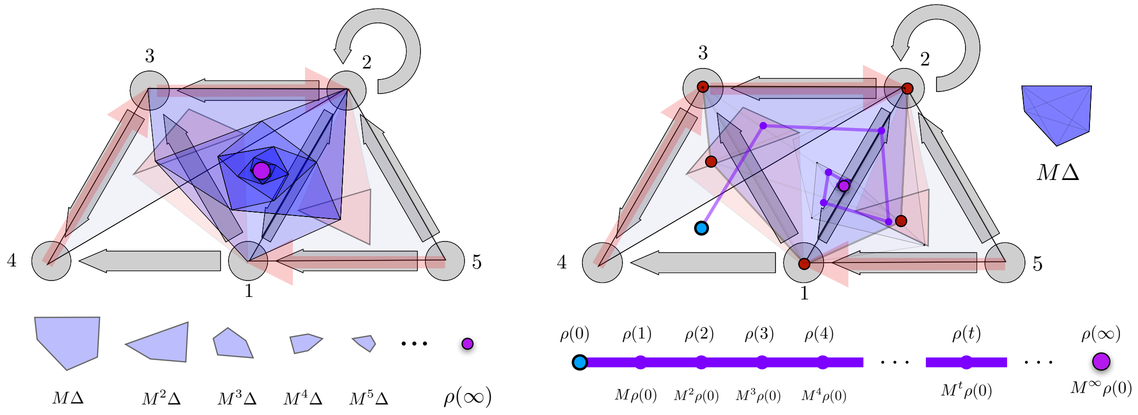

Fig. 5: Trajectory of state distributions converging to the steady state.

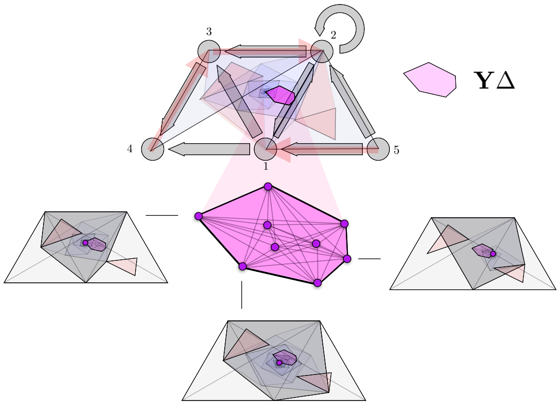

Fig. 6: Illustration of distributions converging to steady state.

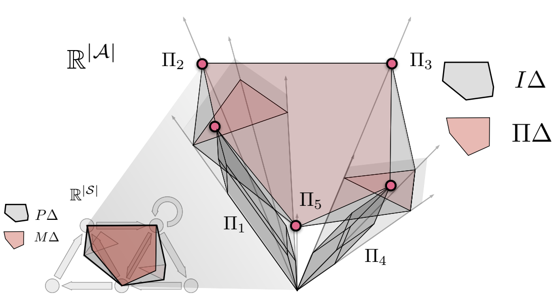

Fig. 7: Action space illustration.

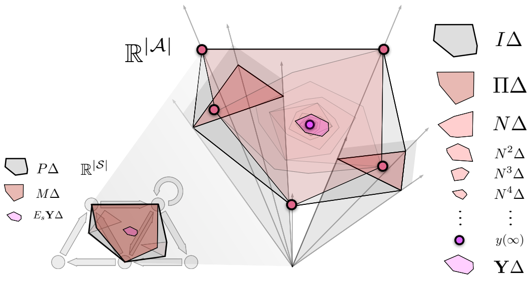

Fig. 8: N^t converging to the steady-state distribution.

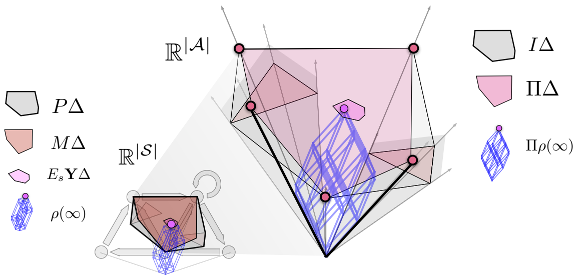

Fig. 9: Steady state distribution hypercube illustration 1.

$$E_s\Pi\rho = P\Pi\rho$$
$$E_s y = Py$$

For given initial distributions, the set of achievable state and state-action distributions over time can be

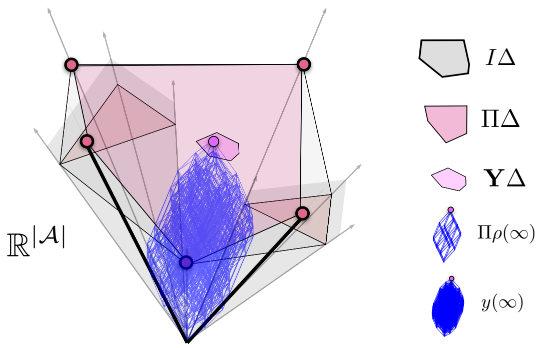

Fig. 10: Steady state distribution hypercube illustration 2.

visualized converging to the steady state distribution shown in Fig. **??** in the state space and **??** in

Matrix: Markov Matrices

$$M = P\Pi \in \mathbb{R}^{|\mathcal{S}|\times|\mathcal{S}|}, \quad N = \Pi P \in \mathbb{R}^{|\mathcal{A}|\times|\mathcal{A}|},$$

Ex:

$$M = P\Pi = \begin{bmatrix} 0 & 0 & 0 & 0.13 & 1 \\ 0.67 & 0 & 0 & 0 & 0 \\ 0.09 & 1 & 0 & 0 & 0 \\ 0.17 & 0 & 1 & 0.24 & 0 \\ 0 & 0 & 0 & 0.7 & 0 \end{bmatrix}$$

$$N = \Pi P = \begin{bmatrix} 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \\ 0.67 & 0.13 & 0.32 & 0.09 & 0.70 & 0.63 & 0 & 0 & 0 \end{bmatrix}$$

Steady-State Eigenvectors:

$$\begin{array}{r|l|l} M: & \substack{\text{Left}\\\text{e-vecs}}: \ \mathbf{1}^{\bar{T}} M = \mathbf{1}^{\bar{T}} & \substack{\text{Right}\\\text{e-vecs}}: \ \rho = M\rho \\ N: & \substack{\text{Left}\\\text{e-vecs}}: \ \mathbf{1}^{\bar{T}} N = \mathbf{1}^{\bar{T}} & \substack{\text{Right}\\\text{e-vecs}}: \ \Pi\rho = M\Pi\rho \end{array}$$

Properties: Column stochastic

B. *Convex Combinations of Distributions*

Since the joint-distribution space is an intersection of an affine space and a convex cone, it is a convex polytope, ie.

$$\{y_k\}_k \in \mathcal{Y}, \quad \Rightarrow \quad y = \sum_k \alpha_k y_k \in \mathcal{Y}$$

for $\sum_k \alpha_k = 1, \ \alpha_k \geq 0$. It is also useful to consider how taking convex combinations of joint-distributions relates to the resulting state-distributions as well as the underlying policies. Steady-state state distributions are also a convex set and taking a convex combination of the joint distributions takes the same convex combination of the state distributions. Specifically, if $\rho_k = E_o y_k$ we have

$$\rho = E_o y = E_o \sum_k \alpha_k y_k = \sum_k \alpha_k E_o y_k = \sum_k \alpha_k \rho_k$$

The relationship between the underlying policies, Π and $\{\Pi_k\}_k$ is more complicated. Specifically since Π is a nonlinear function of y, we have that $\Pi \neq \sum_k \alpha_k \Pi_k$. The specific relationship can be derived from the following

$$\begin{aligned} \Pi &= \mathbf{d}(y) E_o^T \mathbf{d}(\rho)^{-1} \\ &= \mathbf{d}\left(\sum_k \alpha_k y_k\right) E_o^T \mathbf{d}(\rho)^{-1} \\ &= \sum_k \alpha_k \mathbf{d}(y_k) E_o^T \mathbf{d}(\rho_k)^{-1} \mathbf{d}(\rho_k) \mathbf{d}(\rho)^{-1} \\ &= \sum_k \alpha_k \Pi_k \mathbf{d}(\rho_k) \mathbf{d}(\rho)^{-1} \end{aligned}$$

Pulling apart this equation, we have that each element of the policy Π is a convex combination of Π_k but that the coefficients vary from state to state. Specifically within state s we have that

$$\pi_s = \sum_k \left(\frac{\alpha_k \rho_{ks}}{\rho_s}\right) \pi_{ks} = \sum_k \underbrace{\left(\frac{\alpha_k \rho_{ks}}{\sum_k \alpha_k \rho_{ks}}\right)}_{\beta_{ks}} \pi_{ks}$$

Note that this is also a convex combination. However the coefficients are no longer α_k and but rather weighted by the values of the steady state distributions

$$\beta_{ks} = \frac{\alpha_k \rho_{ks}}{\sum_k \alpha_k \rho_{ks}}, \qquad \forall k \tag{8}$$

Note that β_{ks} varies between each state and also that $\beta_{ks} \geq 0$ and $\sum_k \beta_{ks} = 1$.

C. *Convex Combinations of Policies*

We can also consider the affect of taking convex combinations of policies. We first note that taking convex combinations of full policies is not very natural thing to do. A policy by definition is a set of mixed strategies (or convex combinations of actions) at each individual state. For a policy of the form

$$\Pi = \begin{bmatrix} \pi_1 & 0 & \cdots & 0 \\ 0 & \pi_2 & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \pi_{|\mathcal{S}|} \end{bmatrix}$$

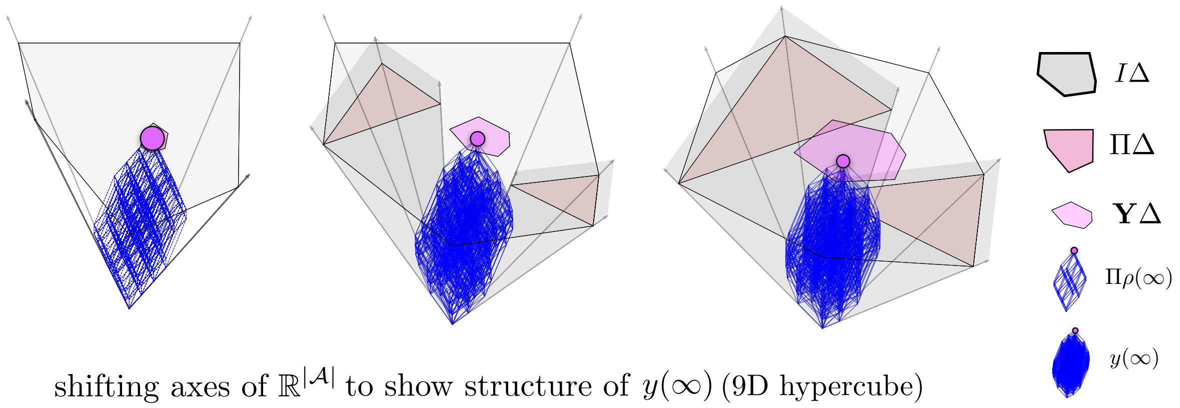
$I\Delta$
$\Pi\Delta$
$\mathbf{Y}\Delta$
$\Pi\rho(\infty)$
$y(\infty)$
shifting axes of $\mathbb{R}^{|\mathcal{A}|}$ to show structure of $y(\infty)$ (9D hypercube)

Fig. 11: Steady state distribution hypercube illustration 2.

each individual vector π_s is a convex combination of pure strategy actions at each state. If we want to take convex combinations of policies, a natural form is to apply a different combination at each point. For a set of policies $\{\Pi_k\}_k$

$$\Pi = \begin{bmatrix} \sum_k \beta_{k1}\pi_{k1} & 0 & \cdots & 0 \\ 0 & \sum_k \beta_{k2}\pi_{k2} & \cdots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \cdots & \sum_k \beta_{k|\mathcal{S}|}\pi_{k|\mathcal{S}|} \end{bmatrix} \tag{9}$$

A convex combination of full policies of the form

$$\Pi = \sum_k \beta_k \Pi_k \tag{10}$$

applies the same combination at each state.

1) Basic Policy Convex Combinations: We start by analyzing steady state distributions of combinations of the form of (10). (We will return to analyzing the steady distributions of policies in the form of (9) which be slightly more involved.) For a combinations of this form, the resulting steady state joint distribution can be derived as follows.

$$\begin{aligned} \Pi\mathbf{d}(\rho) &= \left(\sum_k \beta_k \Pi_k\right)\mathbf{d}(\rho) \\ &= \sum_k \left(\beta_k \Pi_k \mathbf{d}(\rho_k)\mathbf{d}(\rho_k)^{-1}\mathbf{d}(\rho)\right) \\ &= \sum_k \left(\beta_k \Pi_k \mathbf{d}(\rho_k)\mathbf{d}(\rho_k)^{-1}\mathbf{d}(\rho)\right) \end{aligned}$$

Breaking this down by state we have that

$$y_s = \pi_s\rho_s = \sum_k \beta_k \pi_{ks}\rho_{ks}\frac{\rho_s}{\rho_{ks}} = \sum_k y_{ks}\underbrace{\left(\frac{\beta_k\rho_s}{\rho_{ks}}\right)}_{\alpha_{ks}}$$

Note here that the joint distribution at each state is a linear combination of the distributions of the original policies, but this linear combination is state dependent, ie. α_{ks} depends on the state s. Also although $\alpha_{ks} \geq 0$, the set of α's at each state may not be a convex combination, ie. $\sum_k \alpha_{ks} \neq 1$. Without knowing the transition kernel specifically, there is not a way to compare ρ_{ks} and ρ_s and one can easily construct simpe two state MDPs that demonstrate this isn't a convex combination.

2) Combination of Policies that Differ in One State: Before returning to the more natural general combination case (9), we consider the special case, when the policies Π_k are identical at all states except for one. In this case there is actually a one-to-one correspondence between linear combinations of policies and of joint-distributions. To see this without loss of generality, assume the policies are different in state 1 and the same in the other states. This gives that

$$\begin{aligned} \beta_k\Pi_k &= \textbf{blkdiag}\left(\sum_k \beta_k\pi_{k1}, \pi_2, \ldots, \pi_{|\mathcal{S}|}\right) \\ &= \textbf{blkdiag}\left(\sum_k \beta_k\pi_{k1}, \left(\Sigma_k\beta_k''\right)\pi_2, \ldots, \left(\Sigma_k\beta_k^{(|\mathcal{A}|)}\right)\pi_{|\mathcal{S}|}\right) \\ &= \textbf{blkdiag}\left(\sum_k \beta_k\pi_{k1}, \left(\Sigma_k\beta_k''\pi_2\right), \ldots, \left(\Sigma_k\beta_k^{(|\mathcal{A}|)}\pi_{|\mathcal{S}|}\right)\right) \end{aligned}$$

Note that in this second equation we've taken advantage of the fact that the policies are the same in all the states but state 1 to write them as arbitrary convex combinations. The only state that has a fixed convex combination is the first state. Since there is only one state with a fixed combination we can choose the combinations in all states to be consistent with (8) so that the joint distributions have convex combinations given by α_k. To figure out exactly what α_k is for a given set of β_{ks}, we note that the relationship (8) can be written in matrix form as

$$\beta_s = \mathbf{d}(\rho_s)\alpha \frac{1}{\mathbf{1}^T \mathbf{d}(\rho_s)\alpha}$$

where $\beta_s = [\beta_{ks}]_k$, $\alpha = [\alpha_k]_k$ and $\rho_s = [\rho_{ks}]_k$. Solving for $_s$ in terms of α we get

$$\begin{aligned} &\beta_s \mathbf{1}^T \mathbf{d}(\rho_s)\alpha = \mathbf{d}(\rho_s)\alpha \\ \Rightarrow \qquad & 0 = \Big(I - \beta_s \mathbf{1}^T\Big)\mathbf{d}(\rho_s)\alpha \\ \Rightarrow \qquad & \alpha \sim \mathbf{d}(\rho_s)^{-1}\beta_s \end{aligned}$$

This nullspace description reflects the fact that (8) does not restrict the overall scale of the α_k terms. However, mass conservation dictates that if $y = \sum_k \alpha_k y_k$ then $\mathbf{1}^T y = \sum_k \alpha_k \mathbf{1}^T y = \sum_k \alpha_k = 1$ and thus we should scale each α_k to reflect this. This gives

$$\alpha = \frac{\mathbf{d}(\rho_s)^{-1}\beta_s}{\mathbf{1}^T \mathbf{d}(\rho_s)^{-1}\beta_s}$$

or element-wise

$$\alpha_k = \frac{\beta_{ks}/\rho_{ks}}{\sum_k \beta_{ks}/\rho_{ks}} \tag{11}$$

Note the symmetry with (8). We now note that for policies Π_k that differ only in state 1 for a convex combination given by β_k, we have the following construction:

1) Solve for α_k from β_k using (11)
2) Then solve for $\beta_k'', \ldots, \beta_k^{(|\mathcal{A}|)}$ from α_k using (8)

We then have that

$$\begin{aligned} \Pi &= \beta_k \Pi_k \\ &= \textbf{blkdiag}\left(\sum_k \tfrac{\alpha_k \rho_{k1}}{\sum_k \alpha_k \rho_{k1}} \pi_{k1}, , \ldots, \sum_k \tfrac{\alpha_k \rho_{k|\mathcal{S}|}}{\sum_k \alpha_k \rho_{k|\mathcal{S}|}} \pi_{k|\mathcal{S}|}, \right) \end{aligned}$$

Considering each block diagonal element and rearranging gives

$$y_s = \sum_k \alpha_k \underbrace{\rho_{ks}\pi_{ks}}_{y_{ks}} = \pi_s \sum_k \alpha_k \rho_{ks} = \pi_s \rho_s$$

where $y = \sum_k \alpha_k y_k$ and $\rho = \sum_k \alpha_k \rho_k$. We can summarize these insights in the following. If we have a set of policies Π_k that differ in one state only, convex combination in the form $\Pi = \sum_k \beta_k \Pi_k$ results in a convex combination of the joint distributions of the form $y = \sum_k \alpha_k y_k$ where α_k are computed according to (11).

We note that we can plug in the above forms for α_k for the distributions. We produce this here for the state distribution specifically.

Assuming that the policies only differ in state s we have that

$$\begin{aligned} \rho_{s'} &= \sum_k \alpha_k \rho_{ks'} = \sum_k \left(\frac{\beta_{ks}/\rho_{ks}}{\sum_k \beta_{ks}/\rho_{ks}}\right)\rho_{ks'} \quad \forall s' \\ &= \left(\frac{1}{\sum_k \beta_{ks}/\rho_{ks}}\right)\sum_k \left(\beta_{ks}\tfrac{\rho_{ks'}}{\rho_{ks}}\right) \end{aligned}$$

Note that specifically in state s we have

$$\rho_s = 1\Big/\Big(\sum_k \beta_{ks}/\rho_{ks}\Big)$$

3) General Combination of Policies: We now return to the more natural general combination case (9). We note this case is more complicated because in order to define a steady-state distribution, we need the policy to be defined at all the states; and thus the natural way to talk about convex combinations of joint-distributions will involve all possible combinations of strategies at each state.

To be as general as possible, consider index sets of strategies at each state and convex combinations. (Here we will use $n = |\mathcal{S}|$ for notational simplicity.)

$$\begin{aligned} k_1 \in \mathcal{K}_1, \quad \ldots, \quad k_n \in \mathcal{K}_n \\ \beta_1 \in \Delta_{|\mathcal{K}_1|}, \ \ldots, \ \beta_n \in \Delta_{|\mathcal{K}_n|} \end{aligned}$$

We note that if we want to cover all possible policies, we will simply index over the action sets and the convex combinations will be the strategies at each state for a given policy. In this case, the above notation would be replaced with

$$\begin{aligned} \underbrace{a_1}_{k_1} \in \underbrace{\mathcal{A}_1}_{\mathcal{K}_1}, \quad \ldots, \quad \underbrace{a_n}_{k_n} \in \underbrace{\mathcal{A}_n}_{\mathcal{K}_n} \\ \underbrace{\pi_1}_{\beta_1} \in \Delta_{|\mathcal{A}_1|}, \ \ldots, \ \underbrace{\pi_n}_{\beta_n} \in \Delta_{|\mathcal{A}_n|} \end{aligned}$$

Returning to the original notation, we will use the following to indicate the joint and state distributions for all combinations of actions in the index sets as well as for the distributions produced from convex combinations from actions in the sets and mixed combinations

$$\underbrace{\begin{matrix}\rho_s^{k_1,\ldots,k_n}\\ y_s^{k_1,\ldots,k_n}\end{matrix}}_{\text{combinations}}, \qquad \underbrace{\begin{matrix}\rho_s^{\beta_1,\ldots,\beta_n}\\ y_s^{\beta_1,\ldots,\beta_n}\end{matrix}}_{\substack{\text{convex}\\ \text{combinations}}}, \qquad \underbrace{\begin{matrix}\rho_s^{\beta_1\beta_2k_3,\ldots,k_n}\\ y_s^{\beta_1\beta_2k_3,\ldots,k_n}\end{matrix}}_{\substack{\text{mixed}\\ \text{combinations}}},$$

We can then construct the joint distributions from a set of policies using the following steps.

1) **Initialize:** $\left\{ \ \rho_s^{k_1,\ldots,k_n} \ \right\}_{k_1,\ldots,k_n}, \quad \forall s' \in \mathcal{S}$
2) **Loop: for** $s \in \mathcal{S}$
   Given state s:
   For each $k_{s+1},\ldots,k_n$:

$$\rho_{s'}{}^{\beta_1,\ldots,\beta_{s-1},\beta_s,k_{s+1},\ldots,n} = \sum_{k_s} \alpha_{k_s} \rho_{s'}^{\beta_1,\ldots,\beta_{s-1},k_s,k_{s+1},\ldots,n}, \quad \forall s' \in \mathcal{S}$$

   with

$$\alpha_{k_s} = \frac{\beta_{k_s} \Big/ \left(\rho_s^{\beta_1,\ldots,\beta_{s-1},k_s,k_{s+1},\ldots,n}\right)}{\sum_{k_s} \beta_{k_s} \Big/ \left(\rho_s^{\beta_1,\ldots,\beta_{s-1},k_s,k_{s+1},\ldots,n}\right)} \quad \forall k_s \in \mathcal{K}_s$$

Note here the subtle and critical differences in the indexing of each term. Here that at each step we are applying the convex combination $\{\beta_{k_s}\}_{k_s}$ to fix the strategy at state s. Assuming we loop in order through the states, we have to apply this convex combination for all other possible combinations of the remaining indexes (for the unassigned states) $k_{s+1},\ldots,n$. Again, if the index sets are over pure strategies in each state and we're constructing distributions from a policy $(\pi_1,\ldots,\pi_n)$, then $\mathcal{K}_s = \mathcal{A}_s$ and $\{\beta_{k_s}\}_{k_s} = \{\pi_{k_s}\}_{k_s}$. It should also be noted that this is a very inefficient way to construct distributions however it can be elucidating for understanding the structure of the joint distributions related to the policies.

The above process takes repeated convex combinations of distributions. These combinations can be enumerated as

$$\left\{\alpha_{k_1k_2\ldots k_n} = \alpha_{k_1}\alpha_{k_2}\cdots\alpha_{k_n}\right\}_{k_1k_2\ldots k_n}$$

with α_{k_s} computed from the iterative process given above.

D. Joint Distributions as Convex Combinations of Pure Strategies

From the above arguments, we note that we can construct the joint distributions for any policy by taking convex combinations of pure strategy joint distributions. The set of pure stategy policies has size

$$|\mathbf{\Pi}| = \prod_s |\mathcal{A}_s| = |\mathcal{A}_1| \times |\mathcal{A}_2| \times \cdots \times |\mathcal{A}_{|\mathcal{S}|}|$$

and thus the set $\mathcal{Y}$ is a polytope with (up-to) $|\mathbf{\Pi}|$ vertices. We note that this polytope is actually embedded in a $|\mathcal{A}|-|\mathcal{S}|$ dimensional affine space since $\mathcal{Y}$ can be defined by the constraints.

$$\begin{array}{cc} & \leftarrow |\mathcal{A}| \rightarrow \\ \substack{\uparrow \\ |\mathcal{S}|+n_c \\ \downarrow} & \begin{bmatrix} - \ EW \ - \\ - \ \bar{\mathbf{1}}^T \ - \end{bmatrix} y = \begin{bmatrix} 0 \\ 1 \end{bmatrix} \end{array}$$

The dimension of this affine space can be seen since the matrix has a n_c-dimensional left-nullspace and thus has rank $|\mathcal{S}|$ and a $|\mathcal{A}|-\mathcal{S}|$ dimensional right-nullspace (from the rank-nullity theorem). Note in general that $|\mathbf{\Pi}| \gg |\mathcal{A}| - |\mathcal{S}|$ ($\mathcal{Y}$ has many more vertices than the affine space it is embedded in.

One way to enumerate the set of joint distributions is

$$\mathcal{Y} = \left\{ \ y \in \mathbb{R}^{|\mathcal{A}|} \ \middle| \ y = \mathbf{Y}z, \ \mathbf{1}^T z = 1, \ z \geq 0, \ z \in \mathbb{R}^{|\mathbf{\Pi}|} \right\}$$

ie. $y \in \mathbf{Y}\Delta$ where $\mathbf{Y} \in \mathbb{R}^{|\mathcal{A}|\times|\mathbf{\Pi}|}$, is an indicator matrix for the joint distributions

$$\left[\ \mathbf{Y} \ \right]_{a,a_1\cdots a_{|\mathcal{S}|}} = y_a^{a_1\cdots a_{|\mathcal{S}|}};$$

Matrix: Policy Indicator Matrix

$$\mathbf{Y} \in \mathbb{R}^{|\mathcal{A}|\times|\mathbf{\Pi}|}$$

$$\begin{array}{ccc} & \leftarrow \substack{\text{Pure-strategy}\\ \text{policies}} \rightarrow & \\ \mathbf{Y} = & \begin{bmatrix} | & | & & | \\ y_1 & y_2 & \cdots & y_{|\mathbf{\Pi}|} \\ | & | & & | \end{bmatrix} & \substack{\uparrow \\ \text{Actions} \\ \downarrow} \end{array}$$

Ex:

Properties: Column stochastic

The action-edge incidence matrix allows us an interesting interpretation of the feasible set of strategies for the MDP. This relationship is best understood in the edge-flow space illustrated in Fig. 12. We can expand the feasible set to be written as

$$(E_o - E_i)x = 0, \quad x = Wy, \quad \mathbf{1}^T y = 1, \quad y \geq 0$$

The simplex constraint

$$\mathbf{1}^T y = 1, \quad y \geq 0$$

ensures that x lives in the convex hull of the columns of W. Note that since W is a column stochastic matrix, this

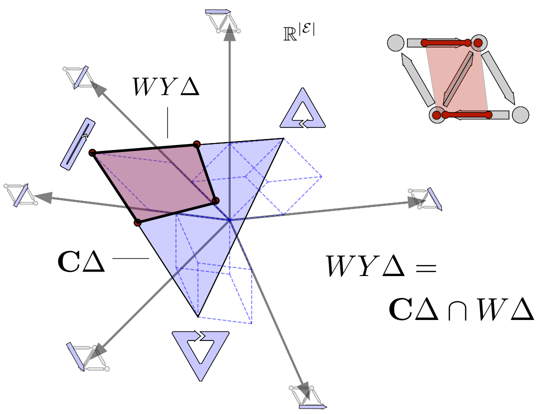

Fig. 12: Illustration of steady-state edge flows as the intersection of the convex-hull $\Delta(W)$ and convex hull of the cycle indicator matrix $\Delta(\mathbf{C})$.

also ensures that $\mathbf{1}^T x = 1,\ x \geq 0$ lives in a simplex in the edge space as well. The incidence matrix constraint

$$(E_o - E_i)x = 0$$

then ensures that x must also live in the nullspace of the incidence matrix and thus be a convex combination of the cycles of the graph. We have then that the edge flow vector x must lie in the intersection of the convex hull of the columns of W and the cycle space of the graph. This relationship is shown in Figure **??**.

asdfasdfasdf

III. Discounted Transition Kernel

We now consider how this optimization problem is modified in the infinite-horizon, discounted-reward case. We introduce a discount factor $0 < \delta < 1$. In the following, we will make use of the following identities (from harmonic series analysis)

$$1 = \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t, \quad (1\text{-}\delta)\delta^t \geq 0,\ \forall t$$

ie., $\{(1\text{-}\delta)\delta^t\}_t$ is a convex combination.

Rather than assume the system is in steady state, we assume that the agent is minimizing the discounted cost

$$\begin{aligned} R_\delta(y) &= \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t r^T y(t) \\ &= r^T \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t y(t) \\ &= r^T y_\delta \end{aligned}$$

for where $y[t]$ is given by the update equation

$$y(t{+}1) = Ny(t), \quad y(0) \in \Delta_{|\mathcal{A}|} \tag{12}$$

and where

$$\begin{aligned} y_\delta &= \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t y(t) \\ &= \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t N^t y(0) \end{aligned}$$

Note that this new *effective joint distribution*, y_δ is indeed a probability distribution, $y_\delta \in I_{|\mathcal{A}|}\Delta$ since it is a convex combination of probability distributions $\{y(t)\}_{t=0}^{\infty}$. Similarly, we could define an *effective state distribution*, $\rho_\delta \in \Delta_{|\mathcal{S}|}$

$$\begin{aligned} \rho_\delta &= \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t \rho(t) \\ &= \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t M^t \rho(0) \end{aligned} \tag{13}$$

The optimal policy from the effective distribution is again given by

$$\pi_a = \tfrac{y_{\delta\, a}}{\sum_a y_{\delta\, a}}$$

One can check that for y_δ given above

$$\begin{aligned} E_s y_\delta &= \delta P y_\delta + (1\text{-}\delta)E_s y(0) \\ E_s y_\delta &= \delta P y_\delta + (1\text{-}\delta)\rho(0) \end{aligned}$$

Indeed,

$$\begin{aligned} E_s y_\delta &= E_s \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t(\Pi P)^t y(0) \\ &= E_s \sum_{t=1}^{\infty}(1\text{-}\delta)\delta^t(\Pi P)^t y(0) + (1\text{-}\delta)E_s y(0) \\ &= \delta E_s \Pi P \sum_{t=1}^{\infty}(1\text{-}\delta)\delta^{t\text{-}1}(\Pi P)^{t\text{-}1} y(0) + (1\text{-}\delta)\rho(0) \\ &= \delta P \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t(\Pi P)^t y(0) + (1\text{-}\delta)\rho(0) \\ &= \delta P y_\delta + (1\text{-}\delta)\rho(0) \end{aligned}$$

In other direction plugging in y_δ with $y_\delta = \Pi\rho_\delta$ gives

$$\begin{aligned} E_s \Pi \rho_\delta &= \delta P \Pi \rho_\delta + (1\text{-}\delta)\rho(0) \\ \rho_\delta &= \delta M \rho_\delta + (\delta)\rho_0 \end{aligned}$$

and iteratively plugging in for ρ_δ gives (13).

Thus we can parametrize the set of y_δ for all possible policies similar to the steady state case in the infinite-horizon average reward setting.

$$Y_\delta = \left\{ \; y_\delta \;\middle|\; E_s y_\delta = \delta P y_\delta + (1\text{-}\delta)\rho(0), \;\; y_\delta \geq 0 \right\} \tag{14}$$

Note the similarities with the steady state version and the additional dependence on the initial state distribution $\rho_0 \in \Delta\big(I_{|\mathcal{S}|}\big)$

Note we could reparametrize (14) as

$$Y_\delta = \left\{ \; y_\delta \;\middle|\; E_s y_\delta = P_\delta y_\delta, \;\; \mathbf{1}^T y_\delta = 1, \;\; y_\delta \geq 0 \right\}$$

where we can define the *equivalent transition kernel* $P_\delta \in [0,1]^{|\mathcal{S}| \times |\mathcal{S}||\mathcal{A}|}$

$$P_\delta = \delta P + (1\text{-}\delta)\rho_0 \mathbf{1}^T$$

Note that the columns of P_δ are simply convex combinations of the columns of P with the initial distribution ρ_0. This can be interpreted as $(1\text{-}\delta)$ of the mass at each state exiting the network at each action and reentering according to the initial distribution. The equivalent transition kernel P_δ is illustrated in Figs. **??** and in detail in **??**.

Again, note that P_δ and Y_δ are very dependent on the initial condition as illustrated in Fig. **??**

We can also define an equivalent edge-action transition kernel.

$$W_\delta = \delta W + (1\text{-}\delta)\Big(I \otimes \rho(0)\mathbf{1}^T\Big)$$

as illustrated in Fig. **??**

Given this definition, we have similar relationships to those defined above.

$$P_\delta = E_i W_\delta, \qquad E_s = E_o W_\delta$$

Matrix: Discounted Transition Kernels

$$P_\delta = \delta P + (1-\delta)\rho_0 \mathbf{1}^T \qquad P_\delta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{A}|}$$
$$W_\delta = \delta W + (1\text{-}\delta)\Big(I \otimes \rho(0)\mathbf{1}^T\Big) \qquad W_\delta \in \mathbb{R}^{|\mathcal{E}| \times |\mathcal{A}|}$$

Properties: Column stochastic

Note, however, in general that $\rho(0)$ will have some positive mass on all states in the network and thus the underlying graph structure will have to be the complete graph. If the underlying graph for the original transition kernel, P, is not the complete graph, it can be expanded to be the complete graph with W having rows of all 0's

From, this equivalent transition kernel we could define equivalent transition matrices that satisfy

$$M_\delta = \delta M + (1\text{-}\delta)\rho_0 \mathbf{1}^T$$
$$N_\delta = \delta N + (1\text{-}\delta)y_0 \mathbf{1}^T$$

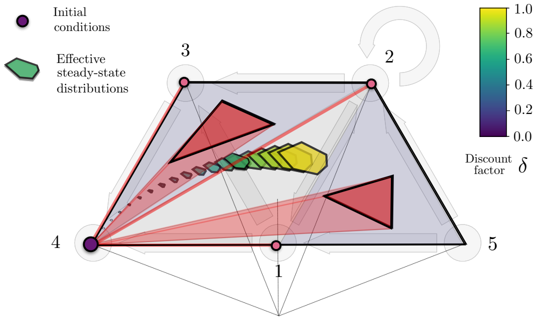

Fig. 13: Discounted effective steady-state distribution for initial condition with transition kernel.

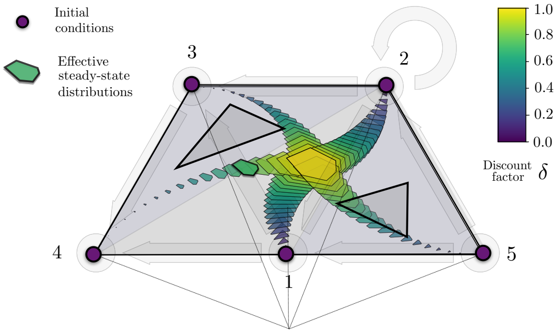

Fig. 14: Discounted effective steady-state distributions for various discount factors and initial conditions.

These transition matrices then satisfy the steady state equations

$$\rho_\delta = M_\delta \rho_\delta = \delta M \rho_\delta + (1\text{-}\delta)\rho_0 \mathbf{1}^T$$
$$y_\delta = N_\delta y_\delta = \delta N y_\delta + (1\text{-}\delta)y_0 \mathbf{1}^T$$

Matrix: Discounted Markov Matrices

$$M_\delta = \delta M + (1\text{-}\delta)\rho_0 \mathbf{1}^T \qquad M_\delta \in \mathbb{R}^{|\mathcal{S}| \times |\mathcal{S}|}$$
$$N_\delta = \delta N + (1\text{-}\delta)y_0 \mathbf{1}^T \qquad N_\delta \in \mathbb{R}^{|\mathcal{A}| \times |\mathcal{A}|}$$

Properties: Column stochastic

Note that

$$\lim_{\delta \to 1} P_\delta = P, \qquad \lim_{\delta \to 0} P_\delta = \rho(0)\mathbf{1}^T$$
$$\lim_{\delta \to 1} Y_\delta = Y, \qquad \lim_{\delta \to 0} Y_\delta = \rho(0)\mathbf{1}^T$$

as illustrated in Fig. **??**

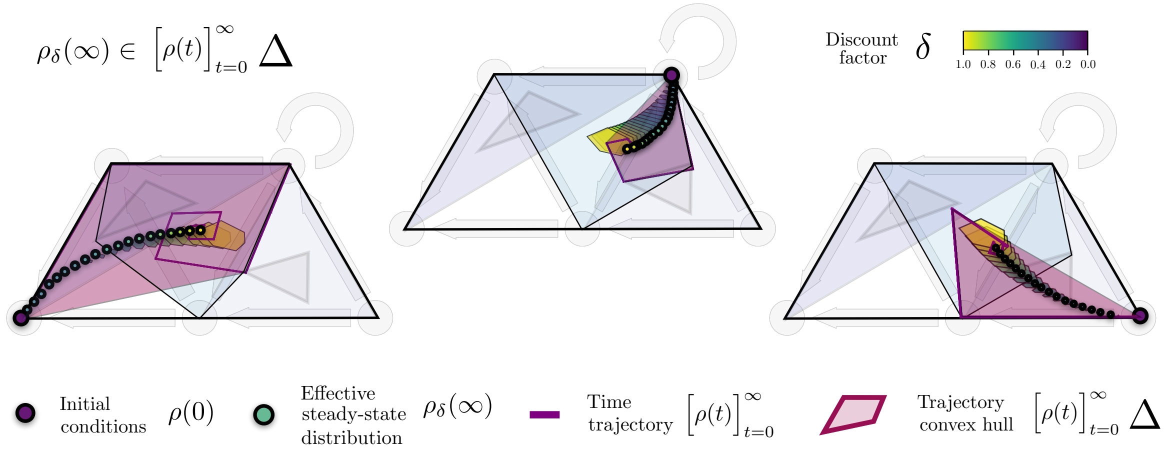
$\rho_\delta(\infty) \in \left[\rho(t)\right]_{t=0}^{\infty} \Delta$
Discount factor δ
1.0 0.8 0.6 0.4 0.2 0.0
Initial conditions $\rho(0)$
Effective steady-state distribution $\rho_\delta(\infty)$
Time trajectory $\left[\rho(t)\right]_{t=0}^{\infty}$
Trajectory convex hull $\left[\rho(t)\right]_{t=0}^{\infty} \Delta$

Fig. 15: Convex hull of trajectories

However, P_δ is a simply a convex combination of P and $\rho(0)\mathbf{1}^T$

$$P_\delta = \delta P + (1\text{-}\delta)\rho(0)\mathbf{1}^T$$

where as Y_δ is convex combination of the whole set trajectory $\{Y(t)\}_{t=0}^{\infty}$.

$$Y_\delta = \sum_{t=0}^{\infty}(1\text{-}\delta)\delta^t Y(t)$$

This convex combination is illustrated (for the pure strategy policies) in Fig. 15.

Matrix: Discounted Policy Indicator

$$\mathbf{Y} \in \mathbb{R}^{|\mathcal{T}||\mathcal{A}| \times |\mathbf{\Pi}|}$$

$$\mathbf{Y} = \begin{array}{c} \leftarrow \text{Pure-strategy policies} \rightarrow \\ \begin{bmatrix} | & | & & | \\ y_1 & y_2 & \cdots & y_{|\mathbf{\Pi}|} \\ | & | & & | \end{bmatrix} \end{array} \begin{array}{c} \uparrow \\ \text{Actions} \\ \downarrow \end{array}$$

Ex: Properties: Column stochastic

IV. Finite-Horizon Transition Kernel

In finite horizon problems, we more specifically model the transients of a system rather than just considering steady state behavior. A finite horizon flow on a graph over a set of time steps $\mathcal{T} = \{0, \cdots T{-}1\}$ can be written as

$$\left\{\mathcal{X}\big(t\big)\right\}_{t\in\mathcal{T}} = \left\{ x(t) \in \mathbb{R}^{|\mathcal{E}|} \;\middle|\; \begin{array}{c} E_o x(0) = \rho(0), \\ E_o x(t{+}1) = E_i x(t), \\ x(t) \geq 0,\ t \in \mathcal{T}, \end{array} \right\}$$

where $\rho(0)$ is the inital mass distribution on the states and $\mathcal{X}\big(t\big)$ represents the reachable $x(t)$'s at time t. When we want to emphasize the emphasis on the initial conditions, we will sometimes use the notation $\mathcal{X}\big(t\big|\rho(0)\big)$.

We may also want to write out this set in matrix form

$$\underbrace{\begin{bmatrix} E_o & 0 & \cdots & 0 & 0 \\ -E_i & E_o & \cdots & 0 & 0 \\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & E_o & 0 \\ 0 & 0 & \cdots & -E_i & E_o \end{bmatrix}}_{\mathbf{E}} \begin{bmatrix} x(0) \\ x(1) \\ \vdots \\ x(T-2) \\ x(T-1) \end{bmatrix} = \begin{bmatrix} \rho(0) \\ 0 \\ \vdots \\ 0 \\ 0 \end{bmatrix}$$

If stochastic transitions are allowed, this constraint incorporates a transition kernel

$$\left\{\mathcal{Y}(t)\right\}_{t\in\mathcal{T}} = \left\{ y(t) \in \mathbb{R}^{|\mathcal{A}|} \;\middle|\; \begin{array}{c} E_s y(0) = \rho(0), \\ E_s y(t{+}1) = P y(t), \\ y(t) \geq 0,\ t \in \mathcal{T} \end{array} \right\}$$

Note that either of these constraints could easily be modified to allow for time varying transitions or mass entering at other points along the time horizon besides $t = 0$.

$$\underbrace{\begin{bmatrix} E_s & 0 & \cdots & 0 & 0\\ -P & E_s & \cdots & 0 & 0\\ \vdots & \vdots & & & \\ 0 & 0 & \cdots & E_s & 0\\ 0 & 0 & \cdots & -P & E_s \end{bmatrix}}_{\mathbf{EW}} \begin{bmatrix} y(0)\\ y(1)\\ \vdots\\ y(T-2)\\ y(T-1) \end{bmatrix} = \begin{bmatrix} \rho(0)\\ 0\\ \vdots\\ 0\\ 0 \end{bmatrix}$$

Note that we may also want to define several of the component matrices separately specifically

$$\begin{aligned} \mathbf{E_o} \in \mathbb{R}^{|\mathcal{S}||\mathcal{T}|\times|\mathcal{E}||\mathcal{T}|}, \quad & \mathbf{E_i} \in \mathbb{R}^{|\mathcal{S}||\mathcal{T}|\times|\mathcal{E}||\mathcal{T}|}, \\ \mathbf{E_s} \in \mathbb{R}^{|\mathcal{S}||\mathcal{T}|\times|\mathcal{A}||\mathcal{T}|}, \quad & \mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{T}|\times|\mathcal{A}||\mathcal{T}|}, \\ & \mathbf{W} \in \mathbb{R}^{|\mathcal{A}||\mathcal{T}|\times|\mathcal{A}||\mathcal{T}|} \end{aligned}$$

separately as

$$\begin{aligned} \mathbf{E_o} = \textbf{blkdiag}(E_o,\dots,E_o), \quad & \mathbf{E_i} = \textbf{blksub1}(E_i,\dots,E_i), \\ \mathbf{E_s} = \textbf{blkdiag}(E_s,\dots,E_s), \quad & \mathbf{P} = \textbf{blksub1}(P,\dots,P), \\ & \mathbf{W} = \textbf{blkdiag}(W,\dots,W) \end{aligned}$$

Similarly, to the infinite horizon cases we have that

$$\mathbf{E} = \mathbf{E_i} - \mathbf{E_o}, \quad \mathbf{E_s} = \mathbf{E_o}\mathbf{W}, \quad \mathbf{P} = \mathbf{E_i}\mathbf{W}$$

And also that each of these matrices is column stochastic

$$\begin{aligned} \mathbf{1}^T\mathbf{E_o} = \mathbf{1}^T, \quad & \mathbf{1}^T\mathbf{E_i} = \mathbf{1}^T, \\ \mathbf{1}^T\mathbf{E_s} = \mathbf{1}^T, \quad \mathbf{1}^T\mathbf{P} = \mathbf{1}^T, \quad & \mathbf{1}^T\mathbf{W} = \mathbf{1}^T, \end{aligned}$$

which again represents mass conservation. We can also define policies in time-dependent setting

$$\mathbf{\Pi} = \textbf{blkdiag}\big(\Pi(0),\dots\Pi(T-1)\big)$$

Note that we then have that

$$(\mathbf{E}_s - \mathbf{P})\mathbf{\Pi}\rho = \underbrace{\begin{bmatrix} I & 0 & \cdots & \cdots & 0\\ -M & I & \cdots & \cdots & 0\\ \vdots & & & & \vdots\\ 0 & 0 & \cdots & -M & I \end{bmatrix}}_{I-\mathbf{M}} \begin{bmatrix} \rho(0)\\ \rho(1)\\ \vdots\\ \rho(T-1) \end{bmatrix} = \begin{bmatrix} \rho_0\\ 0\\ \vdots\\ 0 \end{bmatrix}$$

This affine-space defines the state distribution rollout for a policy $\mathbf{\Pi}$, The right nullspace which defines state distribution rollouts for a given policy $\big(\rho(0),\dots,\rho(T-1)\big)$. Joint-distribution rollout is given by

$$\mathbf{y} = \mathbf{\Pi}\rho = \big(\Pi\rho(0),\dots,\Pi\rho(T-1)\big)$$

A. Distribution Policy Rollouts

Joint distributions in the finite horizon setting can also be written as convex combinations of pure strategy policy rollouts. In the finite-horizon setting the set of pure strategy policies is even larger than in the infinite horizon setting

$$|\mathbf{\Pi}| = \prod_{t\in\mathcal{T}}\prod_{s\in\mathcal{S}} |\mathcal{A}_s|$$

Given a policy $\mathbf{\Pi} = \big(\Pi(0),\dots,\Pi(T-1)\big)$ the state distribution at particular time is given by

$$\begin{aligned} \rho(t) &= M(t-1)\cdots M(1)M(0)\rho_0 \\ &= P\Pi(t-1)\cdots P\Pi(1)P\Pi(0)\rho_0 \end{aligned}$$

Note here that resulting distribution $\rho(t)$ at time t is multi-linear in the policy at each time-step $t' < t$. Thus if we let $\Pi(t') = \sum_k \alpha_k \Pi_k(t')$ for some t' and leave all the other policies at each time step the same then we have

$$\begin{aligned} \rho(t) &= P\Pi(t-1)\cdots P\left(\sum_k \alpha_k \Pi_k(t')\right)\cdots P\Pi(0)\rho_0 \\ &= \sum_k \alpha_k \Big(P\Pi(t-1)\cdots P\Pi_k(t')\cdots P\Pi(0)\rho_0\Big) = \sum_k \alpha_k \rho_k(t) \end{aligned}$$

From a set of possible pure strategy policy rollouts, we can then build up the rollout of any policy as successive convex combinations.

Matrix: Policy Rollout Matrix/Tensor

$$\mathbf{Y} \in \mathbb{R}^{|\mathcal{T}||\mathcal{A}|\times|\mathbf{\Pi}|}$$

$$\mathbf{Y} = \begin{array}{c} \leftarrow \text{Pure-strategy policies} \rightarrow \\ \begin{bmatrix} | & | & & | \\ y_1 & y_2 & \cdots & y_{|\mathbf{\Pi}|} \\ | & | & & | \end{bmatrix} \end{array} \begin{array}{c} \uparrow \\ \text{Times}\times \\ \text{Actions} \\ \downarrow \end{array}$$

Note also that $\mathcal{Y}(t)$ could be represented as $\mathcal{Y}(t) = Y(t)\Delta$ as the convex hull of $Y(t)$ where the columns of $Y(t)$ are computed by enumerating all pure strategy policies up to time t.

Example: Finite horizon flow problems in the state, edge, and action space are illustrated in Fig. **??**.

A policy at each time step can be thought of as a funnel that directs mass from the previous time step to it's updated distribution. These policy evolutions are illustrated in Figs. 16 and **??**.

From various initial condition we can also compute the reachable sets for any policy.

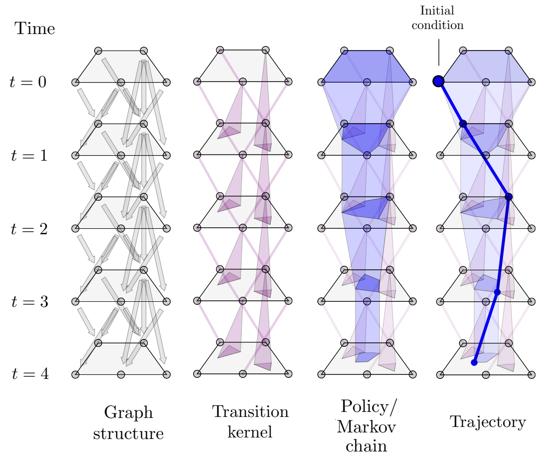

Fig. 16: Finite horizon transition structure

V. Infinite-horizon, average-cost MDP

The infinite horizon, average reward MDP can be thought of selecting the joint steady-state distribution that achieves the highest average reward. Each action has a constant reward assigned to it given by a reward vector $r \in \mathbb{R}^{|\mathcal{A}|}$. We note that if we can break the rewards down into rewards on the states, $r_{\mathcal{S}} \in \mathbb{R}^{|\mathcal{S}|}$, the state actions $r_{\mathcal{A}} \in \mathbb{R}^{|\mathcal{A}|}$ and the edge rewards $r_{\mathcal{E}} \in \mathbb{R}^{|\mathcal{E}|}$ reward we can break r down into various pieces

$$r^T = (r_{\mathcal{S}})^T E_s + (r_{\mathcal{A}})^T + (r_{\mathcal{E}})^T W \tag{15}$$

In the stationary case, we model an agent as optimizing their time average reward. Given Assumption 1 for a particular policy π and the corresponding steady state joint distribution $y \in \Delta_{|\mathcal{A}|}$ the total expected reward is given by $R(y) = r^T y$.

If we enumerate all possible pure strategy policies, in the matrix Y solving for the optimal policy can be done by solving the linear program

$$\left\{ \max_{z} \; r^T Y z \;\middle|\; \mathbf{1}^T z = 1, \; z \geq 0 \right\}$$

Here $z \in \mathbb{R}_+^{|\Gamma|}$ is the mass distribution over the pure strategy policies and $r^T Y$ is vector of rewards for each p.s. policy. Computing the optimal policy in this form can be done graphically by assigning the appropriate rewards r_a to each action and then visualizing the magnitude of $r^T Y_k$ using the method illustrated in Fig. **??**.

This method is shown in Fig. **??** for

$$r^T = \left[1.7, 2.1, 1.3, 2.1, 2.5, 1.7\right]$$

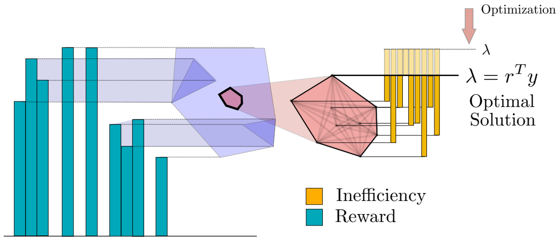

Fig. 17: Illustration of a suboptimal reward distribution for the dual problem. Minimizing λ gives the optimal solution.

The dual formulation of this problem is given by

$$\left\{ \min_{\lambda,\nu} \; \lambda \;\middle|\; \lambda \mathbf{1}^T = r^T Y + \nu^T, \; \nu^T \geq 0 \right\}$$

where the dual variables are $\lambda \in \mathbb{R}$ for the equality constraint and $\nu \in \mathbb{R}_+^{|\Gamma|}$ for the inequality constraint. λ is an upper bound on the maximum average reward for a policy and ν_π is the inefficieny of policy π. Intuitively, solving the dual optimization problem can be pictured as pushing λ as low as possible while keeping $\nu \geq 0$ as shown in Fig. **??**. Optimality is guaranteed by complementary slackness $\nu_\pi z_\pi = 0$, ie. no mass chooses an inefficient policy. Since the overall reward is a linear function of the steady state distribution and the set of steady-state distributions is characterized by a polytope, a pure strategy is optimal (if not uniquely optimal).

A. State Formulation

Alternatively, we can characterize the feasible set of steady state distributions using a more computationally feasible formulation.

$$\left\{ \max_{y} \; r^T y \;\middle|\; E_s y = P y, \; \mathbf{1}^T y = 1, \; y \geq 0 \right\}$$

The dual problem for this formulation is given by

$$\left\{ \min_{\lambda,v,\mu} \; \lambda \;\middle|\; \lambda = r^T + \lambda \mathbf{1}^T + v^T P - v^T E_s + \mu^T, \; \mu^T \geq 0 \right\}$$

At optimum, $\lambda \in \mathbb{R}$ represents the average reward per action. The dual variable $v \in \mathbb{R}^{|\mathcal{S}|}$ represents a value function on each state that encodes how the immediate reward r_a differs from the average reward. $q = v^T P \in \mathbb{R}^{|\mathcal{A}|}$ encodes the reward to go for a particular state-action pair. $\mu \in \mathbb{R}_+^{|\mathcal{A}|}$ represents the inefficiency of any given action. The optimum can be found by minimizing the average reward as much as possible while ensuring that all actions are either optimal or suboptimal, ie. have a positive inefficiency. At optimum, the complementary slackness constraint $y_a \mu_a = 0$ ensures that no suboptimal actions are chosen.

Remark 1. *The constraint above can be rewritten element-wise as*

$$v_s = r_a + q_a - \lambda - \mu_a \tag{16}$$

where $q = v^T P \in \mathbb{R}^{|\mathcal{A}|}$. r_a is the immediate reward for choosing action a from state s; q_a is the expected future reward; μ_a is the inefficiency of that particular action and λ is the equilibrium average reward. v then is a value function on the states that tracks the difference between the immediate reward and the average reward. Equation 16 can be thought of as a version of the Bellman equation.

Lin Program: Primal IH, Average-Reward MDP

Affine representation

$$\max_{y} \; r^T y$$
$$\text{s.t.} \; (E_s - P)y = 0, \; \mathbf{1}^T y = m, \; y \geq 0$$

Vertex representation

$$\max_{y} \; r^T y$$
$$\text{s.t.} \; y = \mathbf{Y}z, \; \mathbf{1}^T z = m, \; z \geq 0$$

Objective & Primal Variables:

Interpret	Joint-distribution	Policy-distribution	Average - reward
Variable	$y \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	$z \in \mathbb{R}_+^{\lvert\Pi\rvert}$	$r^T y$

Constraints & Dual Variables:

Constraint	Interpret	Dual	Interpret
$\mathbf{1}^T y = m$	Total mass conserv.	$\lambda \in \mathbb{R}$	Average reward
$(E_s - P)y = 0$	Local mass conserv.	$v \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	Value function
$y \geq 0$	Mass positive	$\mu \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Inefficiency

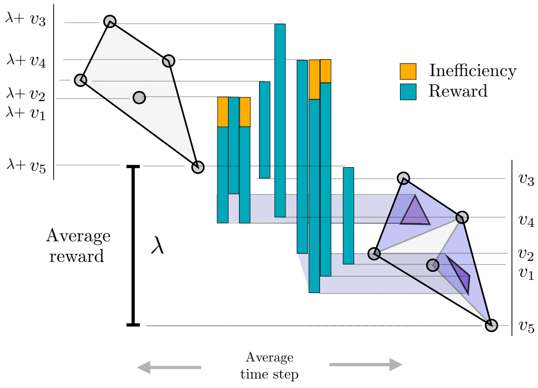

Fig. 18: The optimal solution that satisfies complementary slackness is found by minimizing λ while maintaining $\mu \geq 0$.

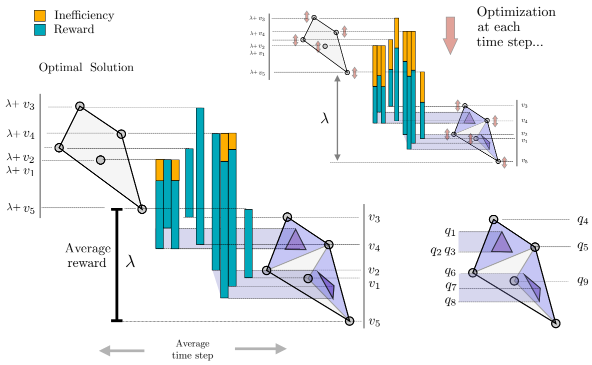

Fig. 19: The optimal solution that satisfies complementary slackness is found by minimizing λ while maintaining $\mu \geq 0$.

Example: An illustration of this dual optimization problem is given in Figs. **??** and **??** for the transition kernel P given before and again

$$r^T = \begin{bmatrix} 1.7 & 2.1 & 1.3 & 2.1 & 2.5 & 1.7 \end{bmatrix}$$

where action $a = 1$ corresponds to state 1, action $a = \{2, 3\}$ correspond to state 2 and action $a = \{4, 5, 6\}$ correspond to state 3. Assigning a height v_s to state s and then super imposing column a of the transition kernel P on the convex hull of the heights gives a way to visualize $q_a = v^T P_a$. Adjusting the base of r_a by q_a allows us to compare the relative heights taking into account the reward-to-go. Any inefficiency with respect to a particular action μ_a is stacked on top of r_a as well. The value of $v_s + \lambda$ for any particular state must be higher than $r_a + \mu_a + q_a$ for any action a associated with that state s. Solving the dual problem involves reducing λ as much as possible while maintaining this restriction. The values of v_s can be adjusted too though both v_s and $v_s + \lambda$ have to differ by λ for all states.

Right multiplying the constraint by a steady state pure strategy joint distribution y (that satisfies $E_s y = Py$, $\mathbf{1}^T y = 1$, $y \geq 0$) gives

$$\lambda \mathbf{1}^T y + v^T E_s y = r^T y + \mu^T y + v^T P y$$
$$\lambda = r^T y + \mu^T y$$

This relationship is illustrated in Fig. **??**.

Lin Program: Dual IH Average-Reward MDP

Affine representation
$$\min_{\lambda,v,\mu} \ \lambda \quad \text{s.t.} \quad \lambda\mathbf{1}^T = v^T(P - E_s) + \mu^T, \ \mu \geq 0$$

Vertex representation
$$\min_{w,\mu} \ \lambda \quad \text{s.t.} \quad \lambda\mathbf{1}^T = r^T\mathbf{R} + \nu^T, \ \nu \geq 0$$

Objective & Primal Variables:

Interpret	Average-reward	Value function	Action inefficiency	Policy inefficiency
Variable	$\lambda \in \mathbb{R}$	$v \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	$\mu \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	$\nu \in \mathbb{R}_+^{\lvert\Pi\rvert}$

Constraints & Dual Variables:

Constraint	Interpret	Dual	Interpret
$\lambda\mathbf{1}^T \geq v^T(P - E_s)$	Ave-reward upper-bnd (actions)	$y \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Joint-distribution
$\lambda\mathbf{1}^T \geq r^T\mathbf{R}$	Ave-reward upper-bnd (policief)	$z \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Policy-distribution

B. *Infinite-Horizon, Discounted-Reward MDP*

If the effective steady state joint distributions are enumerated in the matrix Y_δ, the infinite-horizon, discounted MDP can be formulated as the following linear program.

$$\left\{ \min_{z} \ r^T Y_\delta z \ \middle| \ \mathbf{1}^T z = 1, \ z \geq 0 \right\}$$

A similar technique to the average reward case can be used visualize the solution shown in Fig. **??**.

The reward for a particular pure strategy policy k is

$$\begin{aligned} r^T Y_{\delta k} &= r^T \sum_{t=0}^{\infty} (1\text{-}\delta)\delta^t (P\Pi_k)^t \Pi_k \rho(0) \\ &= \sum_{t=0}^{\infty} (1\text{-}\delta)\delta^t r^T (P\Pi_k)^t \Pi_k \rho(0) \end{aligned}$$

The dual formulation of this problem is given

$$\left\{ \min_{\lambda,\nu} \ \lambda \ \middle| \ \lambda\mathbf{1}^T = r^T Y_\delta + \nu^T, \ \nu^T \geq 0 \right\}$$

where again the dual variable $\lambda \in \mathbb{R}$ for the equality constraint is an upper bound on the maximum discounted reward and $\nu \in \mathbb{R}_+^{|\Gamma|}$ gives the inefficiency of each policy.

Again alternatively, we can characterize the feasible set of effective steady state distributions using a more computationally feasible formulation and the effective transition kernel as

$$\left\{ \max_{y_\delta} \ r^T y_\delta \ \middle| \ E_s y_\delta = P_\delta y_\delta, \ \mathbf{1}^T y_\delta = 1, \ y_\delta \geq 0 \right\}$$

The additional dependence on $\rho(0)$ makes the overall mass conservation constraint redundant and thus we can write this problem as

$$\left\{ \max_{y_\delta} \ r^T y_\delta \ \middle| \ E_s y_\delta = \delta P y_\delta + (1\text{-}\delta)\rho(0)\mathbf{1}^T, \ y_\delta \geq 0 \right\}$$

We will consider the dual problem for this second formulation since it is more commonly used.

$$\left\{ \min_{v,\mu} \ (1\text{-}\delta)v^T\rho(0) \ \middle| \ v^T E_s = r^T + \delta v^T P + \mu^T, \ \mu^T \geq 0 \right\}$$

Lin Program: Primal IH, Disc-Reward MDP

Affine representation
$$\begin{aligned} \max_{y} \quad & r^T y \\ \text{s.t.} \quad & (E_s - \delta P)y = (1 - \delta)\rho(0), \ y \geq 0 \end{aligned}$$

Vertex representation
$$\begin{aligned} \max_{y} \quad & r^T y \\ \text{s.t.} \quad & y = \mathbf{Y}_\delta z, \ \mathbf{1}^T z = 1, \ z \geq 0 \end{aligned}$$

Objective & Primal Variables:

Interpret	Effective joint-distribution	Effective policy-distribution	Discounted-reward
Variable	$y \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	$z \in \mathbb{R}_+^{\lvert\Pi\rvert}$	$r^T y$

Constraints & Dual Variables:

Constraint	Interpret	Dual	Interpret
$\mathbf{1}^T y = m$	Total mass conserv.	$\lambda \in \mathbb{R}$	Discounted reward
$(E_s - P_\delta)y = 0$	Local mass conserv.	$v \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	Effective value function
$y \geq 0$	Mass positive	$\mu \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Inefficiency

The dual variable $v \in \mathbb{R}^{|\mathcal{S}|}$ now represents a discounted reward-to-go on each state. Element-wise the constraint is given by

$$v_s = r_a + \delta q_a + \mu_a$$

where $q = v^T P \in \mathbb{R}^{|\mathcal{A}|}$. r_a is the immediate reward for choosing action a from state s; δq_a is the discounted expected future reward μ_a is the inefficiency of that particular action. Again, Equation **??** is closely related to a discounted Bellman equation. At optimum, the complementary slackness constraint $y_{\delta a}\mu_a = 0$ ensures that no suboptimal actions are chosen. The upper bound on the optimal discounted reward λ is now encoded in the value function v with $\lambda = (1\text{-}\delta)v^T\rho(0)$. The objective is again to minimize this discounted reward while still maintaining positive inefficiency of each action.

This dual program for the discounted problem is illustrated in Fig. **??**

Note that as $\delta \to 0$, the individual rewards are no longer offset by the future reward q_a and only the immediate reward is taken into account.

Right multiplying the constraints by any pure strategy effective distribution $Y_{k\delta}$ for initial distribution $\rho(0)$ gives

$$\begin{aligned} v^T E_s Y_{\delta k} &= r^T Y_{k\delta} + \delta v^T P Y_{k\delta} + \mu^T Y_{k\delta} \\ v^T (E_s - \delta P) Y_{\delta k} &= r^T Y_{k\delta} + \mu^T Y_{k\delta} \\ (1\text{-}\delta) v^T \rho(0) \mathbf{1}^T Y_{\delta k} &= r^T Y_{k\delta} + \mu^T Y_{k\delta} \\ (1\text{-}\delta) v^T \rho(0) &= r^T Y_{k\delta} + \mu^T Y_{k\delta} \end{aligned}$$

This is illustrated for $\delta = 0.6$ and $\delta = 0.0$ in Fig. **??**.

In order to find the optimal policy, it is important that $\rho(0) > 0$ for each element, ie. there is positive mass on each state.

Lin Program: Dual IH, Disc-Reward MDP

Affine represent-ation $\min_{\lambda, v, \mu} \quad \lambda = (1-\delta) v^T \rho(0)$

s.t. $v^T E_s = \delta v^T P + \mu^T, \ \mu \geq 0$

Vertex represent-ation $\min_{w, \mu} \quad \lambda$

s.t. $\lambda \mathbf{1}^T = r^T \mathbf{Y}_\delta + \nu^T, \ \nu \geq 0$

Objective & Primal Variables:

Interpret	Average-reward	Value function	Action inefficiency	Policy inefficiency
Variable	$\lambda \in \mathbb{R}$	$v \in \mathbb{R}^{\lvert\mathcal{S}\rvert}$	$\mu \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	$\nu \in \mathbb{R}_+^{\lvert\Pi\rvert}$

Constraints & Dual Variables:

Constraint	Interpret	Dual	Interpret
$\lambda \mathbf{1}^T \geq v^T (P - E_s)$	Ave-reward upper-bnd (actions)	$y \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Joint-distribution
$\lambda \mathbf{1}^T \geq r^T \mathbf{R}$	Ave-reward upper-bnd (policief)	$z \in \mathbb{R}_+^{\lvert\mathcal{A}\rvert}$	Policy-distribution

VI. Finite-horizon, total-reward MDP

The reward for a finite horizon flow problem is either the total ($\delta = 1$)or discounted reward ($\delta < 1$) summed over time

$$R(y) = \sum_{t \in \mathcal{T}} \delta^t r(t)^T y(t)$$

The finite horizon MDP problem is then given by

$$\left\{ \max_{y(t), t \in \mathcal{T}} \sum_t \delta^t r(t)^T y(t) \;\middle|\; \begin{matrix} E_s y(0) = \rho(0), \\ E_s y(t+1) = P y(t), \\ y(t) \geq 0, \ t \in \mathcal{T} \end{matrix} \right\}$$

The solution to this optimization problem can be visualized as stacking up the rewards received at each time step. This is illustrated in Fig. **??**. With a discount factor, the contribution of the rewards at later time steps is reduced in Fig. **??**.

The illustrations of the rewards for roll outs of policies in the total reward and discounted reward cases are shown in Figs. **??** and **??**

The dual problem for the above optimization problem is given by

$$\left\{ \min_{v(t), \mu(t)} \ v(0)^T \rho(0) \;\middle|\; \begin{matrix} v(t)^T E_s = r(t)^T + \mu(t)^T v(t+1)^T P \\ v(T)^T = r(T)_{\mathcal{S}}^T \\ \mu(t) \geq 0, \ t \in \mathcal{T} \end{matrix} \right\}$$

The dual variables $v(t) \in \mathbb{R}^{|\mathcal{S}|}$ are a value function on the states encoding the reward-to-go. $\mu(T) \in \mathbb{R}_+^{|\mathcal{A}|}$ encodes the ineffiency of particular actions. Elementwise, the constraint is given by

$$v_s(t) = r_a(t) + \mu_a(t) + q_a(t+1), \qquad \forall \ s \in \mathcal{S}, a \in \mathcal{A}$$

where $q(t) = v(t)^T P$ which is the finite horizon Bellman equation. The terminal condition $v(T)^T = r_{\mathcal{S}}(T)^T$ makes the problem directly solvable by dynamic programming. The solution to this problem via dynamic programming is illustrated in Fig. **??**.

In the discounted case, the costs are reduced, Fig. **??**.

At optimum for a given time varying joint distribution that satisfies the constraints, we can right multiply each constraint by $y(t)$ and sum them all to get

$$R = \sum_{t=0}^{T-1} \left(r(t)^T y(t) + \mu(t)^T y(t) \right)$$

and

$$\begin{aligned} R &= \sum_{t=0}^{T-1} v(t)^T E_s y(t) - v(t+1)^T P y(t) \\ &= v(0)^T \rho(0) + \sum_{t=1}^{T-1} v(t)^T E_s y(t) \\ &\quad - \sum_{t=0}^{T-2} v(t+1)^T P y(t) - v(T)^T P y(T-1) \\ &= v(0)^T \rho(0) + \sum_{t=0}^{T-2} v(t+1)^T \left(E_s y(t+1) - P y(t) \right) \\ &\quad - v(T)^T P y(T-1) \\ &= v(0)^T \rho(0) - v(T)^T \rho(T) \end{aligned}$$

The quantity $\sum_t r(t)^T y(t)$ is the reward for that particular policy and $\sum_t \mu(t)^T y(t)$ is the inefficiency of that policy. By complementary slackness, no inefficient policies are chosen at optimum. These roll outs are shown in Figs. **??** and **??**.

References

[1] T. Wang, M. Bowling, and D. Schuurmans, "Dual representations for dynamic programming and reinforcement learning," in *2007 IEEE International Symposium on Approximate Dynamic Programming and Reinforcement Learning*. IEEE, 2007, pp. 44–51.

[2] M. L. Puterman, *Markov decision processes: discrete stochastic dynamic programming*. John Wiley & Sons, 2014.

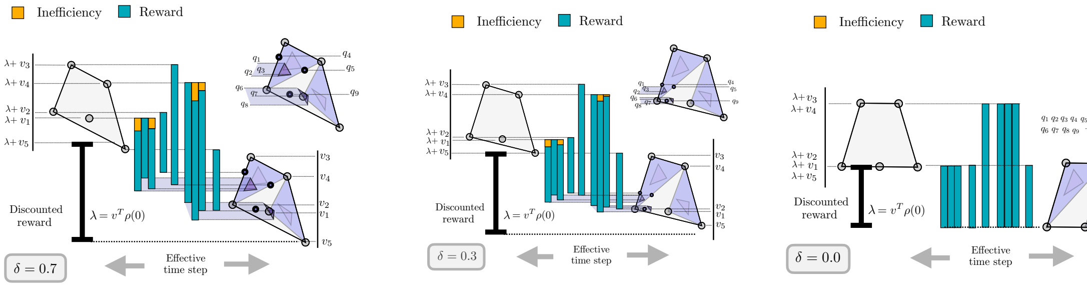

Fig. 20: Dual problem illustration with discounted rewards

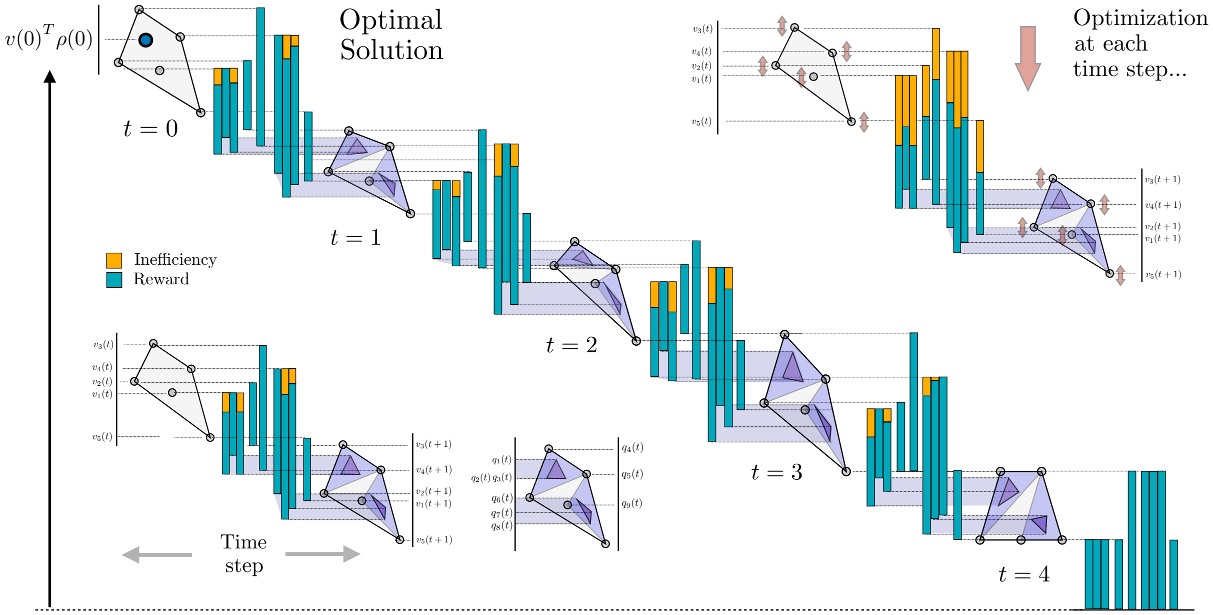

Fig. 21: The optimal solution that satisfies complementary slackness is found by minimizing λ while maintaining $\mu \geq 0$.