Reversible MDP's : Potential Applications?

Alejandro Gomez-Leos

UT Austin, Analysis and Design of Comm. Networks

December 5, 2023

Overview

Overview

1. Formulation [Barto+Sutton, 2018], [Puterman, 2014]
   - An example MDP
   - Basic definitions
   - Avg. reward criterion/policy gain

Overview

1. Formulation [Barto+Sutton, 2018], [Puterman, 2014]
    - An example MDP
    - Basic definitions
    - Avg. reward criterion/policy gain
2. Irreducible/reversible MDPs [Anantharam 2022]

Overview

1. Formulation [Barto+Sutton, 2018], [Puterman, 2014]
   - An example MDP
   - Basic definitions
   - Avg. reward criterion/policy gain
2. Irreducible/reversible MDPs [Anantharam 2022]
3. LP formulation/Poisson's equation, [Ross 1983], [Cogill+Peng, 2013]

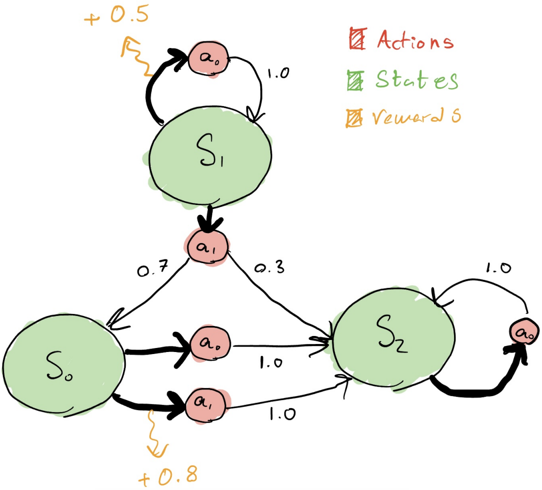
Overview

1. Formulation [Barto+Sutton, 2018], [Puterman, 2014]
    - An example MDP
    - Basic definitions
    - Avg. reward criterion/policy gain
2. Irreducible/reversible MDPs [Anantharam 2022]
3. LP formulation/Poisson's equation, [Ross 1983], [Cogill+Peng, 2013]
4. Policy iteration [Anantharam 2022], [Cogill+Peng, 2013]

Overview

1. Formulation [Barto+Sutton, 2018], [Puterman, 2014]
    - An example MDP
    - Basic definitions
    - Avg. reward criterion/policy gain
2. Irreducible/reversible MDPs [Anantharam 2022]
3. LP formulation/Poisson's equation, [Ross 1983], [Cogill+Peng, 2013]
4. Policy iteration [Anantharam 2022], [Cogill+Peng, 2013]
5. Future Directions

An Example



+ 0.5
Actions
States
rewards
S_1
a_0
1.0
a_1
0.7
0.3
S_0
S_2
a_0
1.0
a_1
1.0
1.0
a_0
+ 0.8

Formulation

Definition

A MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$

Formulation

Definition

A MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$

- $\mathcal{S}$ for states, $\mathcal{A}$ for actions

Formulation

Definition

A MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$

- $\mathcal{S}$ for states, $\mathcal{A}$ for actions
- $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ is state-action-state transition matrix

Formulation

Definition

A MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$

- $\mathcal{S}$ for states, $\mathcal{A}$ for actions
- $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ is state-action-state transition matrix
- $\mathbf{r} \in [0, 1]^{|\mathcal{S}||\mathcal{A}|}$ is reward vector

Formulation

Definition

A MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$

- $\mathcal{S}$ for states, $\mathcal{A}$ for actions
- $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ is state-action-state transition matrix
- $\mathbf{r} \in [0, 1]^{|\mathcal{S}||\mathcal{A}|}$ is reward vector

1. Objects collectively called environment

Formulation

Definition

A MDP is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$

- $\mathcal{S}$ for states, $\mathcal{A}$ for actions
- $\mathbf{P} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}| \times |\mathcal{S}|}$ is state-action-state transition matrix
- $\mathbf{r} \in [0, 1]^{|\mathcal{S}||\mathcal{A}|}$ is reward vector

1. Objects collectively called environment
2. Stochastic process arises by fixing a policy

Policies

Definition

A policy $\pi : \mathcal{S} \rightarrow \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Policies

Definition

A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Notation:

Policies

Definition

A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Notation:

- $\Pi_{\texttt{det}}$: set of deterministic policies for $\mathcal{M}$

Policies

Definition

A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Notation:

- $\Pi_{\texttt{det}}$: set of deterministic policies for $\mathcal{M}$
- $\Pi_{\texttt{rand}}$: set of randomized policies $\mathcal{M}$

Policies

Definition

A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Notation:

- $\Pi_{\texttt{det}}$: set of deterministic policies for $\mathcal{M}$
- $\Pi_{\texttt{rand}}$: set of randomized policies $\mathcal{M}$
- $\Pi_{\texttt{det}} \subseteq \Pi_{\texttt{rand}}$

Policies

Definition

A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Notation:

- $\Pi_{\texttt{det}}$: set of deterministic policies for $\mathcal{M}$
- $\Pi_{\texttt{rand}}$: set of randomized policies $\mathcal{M}$
- $\Pi_{\texttt{det}} \subseteq \Pi_{\texttt{rand}}$
- $\pi(a \mid s)$: prob. of taking $a \in \mathcal{A}$ given in state s

Policies

Definition

A policy $\pi : \mathcal{S} \to \Delta^{|\mathcal{A}|}$ is a conditional distribution over $\mathcal{A}$ for each $s \in \mathcal{S}$.

Notation:

- $\Pi_{\texttt{det}}$: set of deterministic policies for $\mathcal{M}$
- $\Pi_{\texttt{rand}}$: set of randomized policies $\mathcal{M}$
- $\Pi_{\texttt{det}} \subseteq \Pi_{\texttt{rand}}$
- $\pi(a \mid s)$: prob. of taking $a \in \mathcal{A}$ given in state s
- If $\pi \in \Pi_{\texttt{det}}$, $\pi(s)$ is action for state s

Objective

Objective: The agent's goal is to pick a policy π, such that it "accrues maximal reward."

Objective

Objective: The agent's goal is to pick a policy π, such that it "accrues maximal reward."

Several notions of "maximal reward"—different objectives

Formulation

Theorem (Puterman, 2005)

A policy $\pi \in \Pi_{rand}$ is associated with a canonical measure on the sequence of r.v.'s $S_0, A_0, S_1, A_1, \ldots$ for the MDP $\mathcal{M}$.

Formulation

Theorem (Puterman, 2005)

A policy $\pi \in \Pi_{rand}$ is associated with a canonical measure on the sequence of r.v.'s $S_0, A_0, S_1, A_1, \ldots$ for the MDP $\mathcal{M}$.

The performance measure:

- $v^{\pi}_{\mathtt{avg}}(\mathbf{u}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\left(\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t}\right)$, for $\mathbf{u} \in \Delta^{|\mathcal{S}|}$ *aka gain*

Formulation

Theorem (Puterman, 2005)

A policy $\pi \in \Pi_{rand}$ is associated with a canonical measure on the sequence of r.v.'s $S_0, A_0, S_1, A_1, \ldots$ for the MDP $\mathcal{M}$.

The performance measure:

- $v^{\pi}_{\mathtt{avg}}(\mathbf{u}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\big(\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t}\big)$, for $\mathbf{u} \in \Delta^{|\mathcal{S}|}$ *aka gain*

What is a reasonable objective?

Formulation

Theorem (Puterman, 2005)

A policy $\pi \in \Pi_{rand}$ is associated with a canonical measure on the sequence of r.v.'s $S_0, A_0, S_1, A_1, \ldots$ for the MDP $\mathcal{M}$.

The performance measure:

- $v^{\pi}_{\text{avg}}(\mathbf{u}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\left(\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t}\right)$, for $\mathbf{u} \in \Delta^{|\mathcal{S}|}$ *aka gain*

What is a reasonable objective?

- Find π to max $v^{\pi}_{\text{avg}}(\mathbf{u})$ for all $\mathbf{u}$?

Formulation

Theorem (Puterman, 2005)

A policy $\pi \in \Pi_{rand}$ is associated with a canonical measure on the sequence of r.v.'s $S_0, A_0, S_1, A_1, \ldots$ for the MDP $\mathcal{M}$.

The performance measure:

- $v^{\pi}_{\mathrm{avg}}(\mathbf{u}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\left(\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t}\right)$, for $\mathbf{u} \in \Delta^{|\mathcal{S}|}$ *aka gain*

What is a reasonable objective?

- Find π to max $v^{\pi}_{\mathrm{avg}}(\mathbf{u})$ for all $\mathbf{u}$?
- Maybe study $v^{\pi}_{\mathrm{avg}}(\mathbf{u}) = v^{\pi}_{\mathrm{avg}}(\mathbf{u}') \implies$ notion of irreducibility needed.

Irreducible and reversible MDPs

Irreducible and reversible MDPs

- The on-policy transition matrix $\mathbf{P}^\pi$ for chain $(\mathcal{S}, \mathbf{P}^\pi)$ given by $\mathbf{P}^\pi_{s,s'} := \sum_a \mathbf{P}_{(s,a),s'}\, \pi(a \mid s)$

Irreducible and reversible MDPs

- The on-policy transition matrix $\mathbf{P}^\pi$ for chain $(\mathcal{S}, \mathbf{P}^\pi)$ given by $\mathbf{P}^\pi_{s,s'} := \sum_a \mathbf{P}_{(s,a),s'}\, \pi(a \mid s)$

Definition (Irreducible MDP)

If $\forall\, \pi \in \Pi_{\texttt{rand}}$, $(\mathcal{S}, \mathbf{P}^\pi)$ irreducible, then $\mathcal{M}$ is irreducible.

Irreducible and reversible MDPs

- The on-policy transition matrix $\mathbf{P}^\pi$ for chain $(\mathcal{S}, \mathbf{P}^\pi)$ given by $\mathbf{P}^\pi_{s,s'} := \sum_a \mathbf{P}_{(s,a),s'}\, \pi(a \mid s)$

Definition (Irreducible MDP)

If $\forall\, \pi \in \Pi_{\texttt{rand}}$, $(\mathcal{S}, \mathbf{P}^\pi)$ irreducible, then $\mathcal{M}$ is irreducible.

Example:

Irreducible and reversible MDPs

- The on-policy transition matrix $\mathbf{P}^\pi$ for chain $(\mathcal{S}, \mathbf{P}^\pi)$ given by $\mathbf{P}^\pi_{s,s'} := \sum_a \mathbf{P}_{(s,a),s'} \, \pi(a \mid s)$

Definition (Irreducible MDP)

If $\forall\, \pi \in \Pi_{\mathtt{rand}}$, $(\mathcal{S}, \mathbf{P}^\pi)$ irreducible, then $\mathcal{M}$ is irreducible.

Example: Any MDP with $\mathbf{P}_{(s,a),s'} > 0$ for all (s, a, s')

Irreducible and reversible MDPs

- The on-policy transition matrix $\mathbf{P}^{\pi}$ for chain $(\mathcal{S}, \mathbf{P}^{\pi})$ given by $\mathbf{P}^{\pi}_{s,s'} := \sum_a \mathbf{P}_{(s,a),s'} \, \pi(a \mid s)$

Definition (Irreducible MDP)

If $\forall \, \pi \in \Pi_{\mathtt{rand}}$, $(\mathcal{S}, \mathbf{P}^{\pi})$ irreducible, then $\mathcal{M}$ is irreducible.

Example: Any MDP with $\mathbf{P}_{(s,a),s'} > 0$ for all (s, a, s')

Definition (Reversible MDP, Cogill+Peng, 2013, Anantharam 2022)

If $\forall \, \pi \in \Pi_{\mathtt{rand}}$, $(\mathcal{S}, \mathbf{P}^{\pi})$ irreducible and reversible, then $\mathcal{M}$ is reversible.

Irreducible and reversible MDPs

- The on-policy transition matrix $\mathbf{P}^\pi$ for chain $(\mathcal{S}, \mathbf{P}^\pi)$ given by $\mathbf{P}^\pi_{s,s'} := \sum_a \mathbf{P}_{(s,a),s'}\,\pi(a \mid s)$

Definition (Irreducible MDP)

If $\forall\, \pi \in \Pi_{\mathtt{rand}}$, $(\mathcal{S}, \mathbf{P}^\pi)$ irreducible, then $\mathcal{M}$ is irreducible.

Example: Any MDP with $\mathbf{P}_{(s,a),s'} > 0$ for all (s, a, s')

Definition (Reversible MDP, Cogill+Peng, 2013, Anantharam 2022)

If $\forall\, \pi \in \Pi_{\mathtt{rand}}$, $(\mathcal{S}, \mathbf{P}^\pi)$ irreducible and reversible, then $\mathcal{M}$ is reversible.

Note: Could replace $\Pi_{\mathtt{rand}}$ by $\Pi_{\mathtt{det}}$ in def's above.

Reversible MDP Example

Theorem (Anantharam, 2022)

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let* $\mathcal{S}$, $\mathcal{A}$ *be any finite sets*
2. *Let* $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ *any simple connected graph, with*

$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*
2. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s'), *where* $w_s := \sum_{s'} w_{s,s'}$

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*
2. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s'), *where* $w_s := \sum_{s'} w_{s,s'}$
4. *Let* $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let* $\mathcal{S}$, $\mathcal{A}$ *be any finite sets*
2. *Let* $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ *any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s')*, where* $w_s := \sum_{s'} w_{s,s'}$
4. *Let* $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$
5. $\mathbf{P}_{(s,a),s} := 1 - \rho(s, a), \quad \mathbf{P}_{(s,a),s'} := \rho(s, a)\, \mathbf{Q}_{s,s'} \quad \mathbf{r}_{s,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*
2. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s'), *where* $w_s := \sum_{s'} w_{s,s'}$
4. *Let* $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$
5. $\mathbf{P}_{(s,a),s} := 1 - \rho(s, a), \quad \mathbf{P}_{(s,a),s'} := \rho(s, a)\, \mathbf{Q}_{s,s'} \quad \mathbf{r}_{s,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$\mathcal{M} := \left(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r}\right)$ *is a reversible MDP.*

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*
2. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s'), *where* $w_s := \sum_{s'} w_{s,s'}$
4. *Let* $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$
5. $\mathbf{P}_{(s,a),s} := 1 - \rho(s, a), \quad \mathbf{P}_{(s,a),s'} := \rho(s, a)\, \mathbf{Q}_{s,s'} \quad \mathbf{r}_{s,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ *is a reversible MDP.*

Interpretation:

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*
2. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s'), *where* $w_s := \sum_{s'} w_{s,s'}$
4. *Let* $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$
5. $\mathbf{P}_{(s,a),s} := 1 - \rho(s, a), \quad \mathbf{P}_{(s,a),s'} := \rho(s, a)\, \mathbf{Q}_{s,s'} \quad \mathbf{r}_{s,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$\mathcal{M} := (\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r})$ *is a reversible MDP.*

Interpretation:

- like *lazy weighted random walk*, with variable laziness via ρ

Reversible MDP Example

Theorem (Anantharam, 2022)

1. *Let $\mathcal{S}$, $\mathcal{A}$ be any finite sets*
2. *Let $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ any simple connected graph, with*
$$(s, s') \in \mathcal{E} \iff \exists\, w_{s,s'} = w_{s',s} > 0$$
3. $\mathbf{Q}_{s,s'} = \frac{w_{s,s'}}{w_s}$ *for all* (s, s'), *where* $w_s := \sum_{s'} w_{s,s'}$
4. *Let* $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$
5. $\mathbf{P}_{(s,a),s} := 1 - \rho(s, a), \quad \mathbf{P}_{(s,a),s'} := \rho(s, a)\, \mathbf{Q}_{s,s'} \quad \mathbf{r}_{s,a} \in \mathbb{R}^{|\mathcal{S}||\mathcal{A}|}$

$\mathcal{M} := \left(\mathcal{S}, \mathcal{A}, \mathbf{P}, \mathbf{r}\right)$ *is a reversible MDP.*

Interpretation:

- like *lazy weighted random walk*, with variable laziness via ρ
- actions can control laziness, but not totally

A more general characterization

- This construction exhibits a reversible MDP

A more general characterization

- This construction exhibits a reversible MDP
- Conversely, do all reversible MDPs have that structure?

A more general characterization

- This construction exhibits a reversible MDP
- Conversely, do all reversible MDPs have that structure? Sort of!

A more general characterization

- This construction exhibits a reversible MDP
- Conversely, do all reversible MDPs have that structure? Sort of!

Lemma (Anatharam, '22)

Consider reversible MDP $\mathcal{M}$. There exists a simple connected graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ s.t. for all $a \in \mathcal{A}$ and $s \neq s'$, $P_{(s,a),s'} > 0$ iff $(s, s') \in E$.

A more general characterization

- This construction exhibits a reversible MDP
- Conversely, do all reversible MDPs have that structure? Sort of!

Lemma (Anatharam, '22)

Consider reversible MDP $\mathcal{M}$. There exists a simple connected graph $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ s.t. for all $a \in \mathcal{A}$ and $s \neq s'$, $P_{(s,a),s'} > 0$ iff $(s, s') \in E$.

Theorem ("bi-connection theorem" Anatharam, '22)

Let $\mathcal{M}$ be a reversible MDP, and $\mathcal{G} = (\mathcal{S}, \mathcal{E})$ its canonical graph. If $\mathcal{G}$ is bi-connected, then there exists a irr. and reversible $\mathbf{Q}$ and a function $\rho : \mathcal{S} \times \mathcal{A} :\to (0, 1]$ such that for each $a \in \mathcal{A}$,

- $\mathbf{Q}_{s,s'} > 0$ *iff* $(s, s') \in \mathcal{E}$
- $\mathbf{P}_{(s,a),s} := 1 - \rho(s, a), \quad \mathbf{P}_{(s,a),s'} := \rho(s, a)\mathbf{Q}_{s,s'}$

Optimization

Back to optimization.

Back to optimization.
Suppose $\mathcal{M}$ is irreducible, recall:

Optimization

Back to optimization.

Suppose $\mathcal{M}$ is irreducible, recall:

$$v^{\pi}_{\mathrm{avg}}(\mathbf{u}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\left(\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t} \right), \text{ for } \mathbf{u} \in \Delta^{|\mathcal{S}|}$$

Optimization

Back to optimization.
Suppose $\mathcal{M}$ is irreducible, recall:

$$v^{\pi}_{\mathrm{avg}}(\mathbf{u}) := \lim_{T\to\infty} \frac{1}{T}\mathbb{E}_{\pi}\big(\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t}\big),\text{ for } \mathbf{u} \in \Delta^{|\mathcal{S}|}$$

For any π, $(\mathcal{S}, \mathbf{P}^{\pi})$ is irreducible, so sum converges to same quantity regardless of $\mathbf{u}$. One can show

$$v^{\pi}_{\mathrm{avg}} = \sum_{s,a} \mu^{\pi}(s)\,\pi(a \mid s)\,\mathbf{r}_{s,a}$$

Optimization

Back to optimization.
Suppose $\mathcal{M}$ is irreducible, recall:

$$v^{\pi}_{\mathrm{avg}}(\mathbf{u}) := \lim_{T\to\infty} \tfrac{1}{T}\mathbb{E}_{\pi}\big(\textstyle\sum_{t=0}^{T-1} \mathbf{r}_{S_t,A_t}\big), \text{ for } \mathbf{u} \in \Delta^{|\mathcal{S}|}$$

For any π, $(\mathcal{S}, \mathbf{P}^{\pi})$ is irreducible, so sum converges to same quantity regardless of $\mathbf{u}$. One can show

$$v^{\pi}_{\mathrm{avg}} = \textstyle\sum_{s,a} \mu^{\pi}(s)\,\pi(a \mid s)\,\mathbf{r}_{s,a}$$

How to pick π to maximize?

Primal Formulation

$$v_{\text{avg}}^{\pi} = \sum_{s,a} \underbrace{\mu^{\pi}(s)\,\pi(a \mid s)}_{:=\phi^{\pi}(s,a)} \mathbf{r}_{s,a}$$

Primal Formulation

$$v^{\pi}_{\text{avg}} = \sum_{s,a} \underbrace{\mu^{\pi}(s)\,\pi(a \mid s)}_{:=\phi^{\pi}(s,a)} \mathbf{r}_{s,a}$$

Solve the following LP OPT:

$$\begin{aligned}
\max_{\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \quad & \sum_{s,a} \phi(s,a)\,\mathbf{r}_{s,a} \\
\text{s.t.} \quad & \phi(s,a) \geq 0 \quad \forall s,a \\
& \sum_{s,a} \phi(s,a) = 1 \\
& \sum_{s,a} \phi(s,a) = \sum_{s',a'} \mathbf{P}_{(s',a'),s}\,\phi(s',a') \quad \forall s
\end{aligned}$$

Primal Formulation

$$v^\pi_{\text{avg}} = \sum_{s,a} \underbrace{\mu^\pi(s)\,\pi(a \mid s)}_{:=\phi^\pi(s,a)} \mathbf{r}_{s,a}$$

Solve the following LP OPT:

$$\begin{aligned}
\max_{\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \quad & \sum_{s,a} \phi(s,a)\,\mathbf{r}_{s,a} \\
\text{s.t.} \quad & \phi(s,a) \geq 0 \quad \forall s,a \\
& \sum_{s,a} \phi(s,a) = 1 \\
& \sum_{s,a} \phi(s,a) = \sum_{s',a'} \mathbf{P}_{(s',a'),s}\,\phi(s',a') \quad \forall s
\end{aligned}$$

Note: feasible thanks to irreducibility of $\mathcal{M}$—but, does there exists an optimal solution?

Primal Formulation

$$v^{\pi}_{\text{avg}} = \sum_{s,a} \underbrace{\mu^{\pi}(s)\,\pi(a \mid s)}_{:=\phi^{\pi}(s,a)} \mathbf{r}_{s,a}$$

Solve the following LP $\texttt{OPT}$:

$$\begin{aligned}
\max_{\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \quad & \sum_{s,a} \phi(s,a)\,\mathbf{r}_{s,a} \\
\text{s.t.} \quad & \phi(s,a) \geq 0 \quad \forall s,a \\
& \sum_{s,a} \phi(s,a) = 1 \\
& \sum_{s,a} \phi(s,a) = \sum_{s',a'} \mathbf{P}_{(s',a'),s}\,\phi(s',a') \quad \forall s
\end{aligned}$$

Note: feasible thanks to irreducibility of $\mathcal{M}$—but, does there exists an optimal solution? Yes!

Dual LP of OPT

$$
\begin{aligned}
\max_{\phi:\mathcal{S}\times\mathcal{A}\to\mathbb{R}} \quad & \sum_{s,a} \phi(s,a)\, \mathbf{r}_{s,a} \\
\text{s.t.} \quad & \phi(s,a) \geq 0 \quad \forall s,a \\
& \sum_{s,a} \phi(s,a) = 1 \\
& \sum_{s,a} \phi(s,a) = \sum_{s',a'} P_{(s',a'),s} \phi(s',a') \quad \forall s
\end{aligned}
$$

OPT admits the dual LP:

$$
\begin{aligned}
\min_{h:\mathcal{S}\to\mathbb{R},\ C\in\mathbb{R}} \quad & C \\
\text{s.t.} \quad & \mathbf{r}_{s,a} + \sum_{s'} h(s')\, \mathbf{P}_{(s,a),s'} \leq h(s) + C \quad \forall s,a
\end{aligned}
$$

Existence of optimal det. policy

Claim: We can find a$\pi \in \Pi_{\mathtt{det}}$ such that $\phi^{\pi} = \phi^{*}$ for `OPT`.

Existence of optimal det. policy

Claim: We can find a$\pi \in \Pi_{\mathtt{det}}$ such that $\phi^{\pi} = \phi^{*}$ for `OPT`.

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that

$$C + h(s) = \max_{a} \left(\mathbf{r}_{s,a} + \sum_{s'} \mathbf{P}_{(s,a),s'} \, h(s') \right) \quad \forall \, s \in \mathcal{S}$$

then there exists an optimal $\pi \in \Pi_{det}$, where $v^{\pi}_{avg} = C$.

Existence of optimal det. policy

Claim: We can find a$\pi \in \Pi_{\mathtt{det}}$ such that $\phi^\pi = \phi^*$ for $\mathtt{OPT}$.

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that

$$C + h(s) = \max_a \left(\mathbf{r}_{s,a} + \sum_{s'} \mathbf{P}_{(s,a),s'}\, h(s')\right) \quad \forall\, s \in \mathcal{S}$$

then there exists an optimal $\pi \in \Pi_{det}$, where $v^\pi_{avg} = C$.

Notation: For more compact formulas, define

$$\psi^h(s,a) := \mathbf{r}_{s,a} + \textstyle\sum_{s'} \mathbf{P}_{(s,a),s'}\, h(s')$$

for each given h and pair (s,a)

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that $C + h(s) = \max_a \psi^h(s, a)$ for all $s \in \mathcal{S}$, then there exists an optimal $\pi \in \Pi_{det}$, where $v^{\pi}_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\mathtt{det}}$ such that $\phi^{\pi} = \phi^{*}$ for `OPT`

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that $C + h(s) = \max_a \psi^h(s, a)$ for all $s \in \mathcal{S}$, then there exists an optimal $\pi \in \Pi_{det}$, where $v^\pi_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\mathtt{det}}$ such that $\phi^\pi = \phi^*$ for OPT

Proof.

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that $C + h(s) = \max_a \psi^h(s, a)$ for all $s \in \mathcal{S}$, then there exists an optimal $\pi \in \Pi_{det}$, where $v^{\pi}_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\mathtt{det}}$ such that $\phi^{\pi} = \phi^*$ for `OPT`

Proof.

Pick (h^*, C^*) as dual optimal sol'ns. Goal is to exhibit they satisfy lemma.

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that $C + h(s) = \max_a \psi^h(s, a)$ for all $s \in \mathcal{S}$, then there exists an optimal $\pi \in \Pi_{det}$, where $v^{\pi}_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\mathtt{det}}$ such that $\phi^{\pi} = \phi^*$ for `OPT`

Proof.

Pick (h^*, C^*) as dual optimal sol'ns. Goal is to exhibit they satisfy lemma.

For each s, let $a_s := \arg\max_a \psi^{h^*}(s, a)$.

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, *if there is a bounded real function* $h : \mathcal{S} \to \mathbb{R}$, *and a constant* C, *such that* $C + h(s) = \max_a \psi^h(s, a)$ *for all* $s \in \mathcal{S}$, *then there exists an optimal* $\pi \in \Pi_{det}$, *where* $v^{\pi}_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\mathtt{det}}$ such that $\phi^{\pi} = \phi^*$ for `OPT`

Proof.

Pick (h^*, C^*) as dual optimal sol'ns. Goal is to exhibit they satisfy lemma.

For each s, let $a_s := \arg\max_a \psi^{h^*}(s, a)$.
BWOC there exists $\tilde{s}$ s.t. $\psi^{h^*}(\tilde{s}, a_{\tilde{s}}) < h^*(\tilde{s}) + C^*$.

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that $C + h(s) = \max_a \psi^h(s, a)$ for all $s \in \mathcal{S}$, then there exists an optimal $\pi \in \Pi_{det}$, where $v^{\pi}_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\mathtt{det}}$ such that $\phi^\pi = \phi^*$ for $\mathtt{OPT}$

Proof.

Pick (h^*, C^*) as dual optimal sol'ns. Goal is to exhibit they satisfy lemma.

For each s, let $a_s := \arg\max_a \psi^{h^*}(s, a)$.
BWOC there exists $\tilde{s}$ s.t. $\psi^{h^*}(\tilde{s}, a_{\tilde{s}}) < h^*(\tilde{s}) + C^*$.
By C.S. $\phi^*(\tilde{s}, a_{\tilde{s}}) = 0$, so $\exists\, a' : \phi^*(\tilde{s}, a') > 0 \implies \psi^{h^*}(\tilde{s}, a') = h^*(\tilde{s}) + C^*$.

Continuing

Lemma (S.M. Ross, 1983)

For any irreducible MDP $\mathcal{M}$, if there is a bounded real function $h : \mathcal{S} \to \mathbb{R}$, and a constant C, such that $C + h(s) = \max_a \psi^h(s, a)$ for all $s \in \mathcal{S}$, then there exists an optimal $\pi \in \Pi_{det}$, where $v^{\pi}_{avg} = C$.

Claim: There exists a policy $\pi \in \Pi_{\texttt{det}}$ such that $\phi^\pi = \phi^*$ for $\texttt{OPT}$

Proof.

Pick (h^*, C^*) as dual optimal sol'ns. Goal is to exhibit they satisfy lemma.

For each s, let $a_s := \arg\max_a \psi^{h^*}(s, a)$.
BWOC there exists $\tilde{s}$ s.t. $\psi^{h^*}(\tilde{s}, a_{\tilde{s}}) < h^*(\tilde{s}) + C^*$.
By C.S. $\phi^*(\tilde{s}, a_{\tilde{s}}) = 0$, so $\exists\, a' : \phi^*(\tilde{s}, a') > 0 \implies \psi^{h^*}(\tilde{s}, a') = h^*(\tilde{s}) + C^*$.

But $h^*(\tilde{s}) + C^* = \psi^{h^*}(\tilde{s}, a') \leq \psi^{h^*}(\tilde{s}, a_{\tilde{s}}) < h^*(\tilde{s}) + C^*$, contradiction.

Dual LP of OPT properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$

Dual LP of OPT properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\mathtt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Dual LP of $\texttt{OPT}$ properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Take $\phi^* = \phi^{\pi *}$ (this is OK because $\phi^{\pi *}$ is primal-optimal).

Dual LP of $\texttt{OPT}$ properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Take $\phi^* = \phi^{\pi*}$ (this is OK because $\phi^{\pi*}$ is primal-optimal). Fix an $s \in \mathcal{S}$.

Dual LP of OPT properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\mathtt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Take $\phi^* = \phi^{\pi *}$ (this is OK because $\phi^{\pi *}$ is primal-optimal). Fix an $s \in \mathcal{S}$. The following holds iff (h^*, C^*) are dual-optimal by S.D.

Dual LP of OPT properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Take $\phi^* = \phi^{\pi *}$ (this is OK because $\phi^{\pi *}$ is primal-optimal). Fix an $s \in \mathcal{S}$. The following holds iff (h^*, C^*) are dual-optimal by S.D.

1. If $a \neq \pi^*(s)$, then $\phi^*(s,a) = 0 \implies \psi^{h^*}(s,a) < h^*(s) + C^*$

Dual LP of $\texttt{OPT}$ properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Take $\phi^* = \phi^{\pi *}$ (this is OK because $\phi^{\pi *}$ is primal-optimal). Fix an $s \in \mathcal{S}$. The following holds iff (h^*, C^*) are dual-optimal by S.D.

1. If $a \neq \pi^*(s)$, then $\phi^*(s,a) = 0 \implies \psi^{h^*}(s,a) < h^*(s) + C^*$
2. If $a = \pi^*(s)$, then $\phi^*(s,a) > 0 \implies \psi^{h^*}(s,a) = h^*(s) + C^*$

Dual LP of OPT properties

Facts:

1. $\exists\, \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Proof.

Take $\phi^* = \phi^{\pi *}$ (this is OK because $\phi^{\pi *}$ is primal-optimal). Fix an $s \in \mathcal{S}$. The following holds iff (h^*, C^*) are dual-optimal by S.D.

1. If $a \neq \pi^*(s)$, then $\phi^*(s,a) = 0 \implies \psi^{h^*}(s,a) < h^*(s) + C^*$
2. If $a = \pi^*(s)$, then $\phi^*(s,a) > 0 \implies \psi^{h^*}(s,a) = h^*(s) + C^*$

Hence $\max_a \psi^{h^*}(s,a) = h^*(s) + C^*$.

Dual LP of OPT properties

Facts:

1. $\exists \ \pi^* \in \Pi_{\mathtt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s, a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s, a) = h^*(s) + C$ for all s.

Dual LP of OPT properties

Facts:

1. $\exists \quad \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Policy Iteration:

Dual LP of OPT properties

Facts:

1. $\exists \;\; \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi *} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Policy Iteration: Input initial $\pi_0 \in \Pi_{\texttt{det}}$, for $k = 0, 1, 2, \ldots$

Dual LP of OPT properties

Facts:

1. $\exists \quad \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Policy Iteration: Input initial $\pi_0 \in \Pi_{\texttt{det}}$, for $k = 0, 1, 2, \ldots$

- Solve *Poisson's eq'n*

$$\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$$

  for (h^{π_k}, C^{π_k}) (system of $|\mathcal{S}|$ eq'ns in $|\mathcal{S}| + 1$ unknowns)

Dual LP of OPT properties

Facts:

1. $\exists \quad \pi^* \in \Pi_{\mathtt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Policy Iteration: Input initial $\pi_0 \in \Pi_{\mathtt{det}}$, for $k = 0, 1, 2, \ldots$

- Solve *Poisson's eq'n*

$$\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$$

  for (h^{π_k}, C^{π_k}) (system of $|\mathcal{S}|$ eq'ns in $|\mathcal{S}| + 1$ unknowns)

- Check if

$$\max_a \psi^{h^{\pi_k}}(s, a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$$

Dual LP of OPT properties

Facts:

1. $\exists \quad \pi^* \in \Pi_{\mathtt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Policy Iteration: Input initial $\pi_0 \in \Pi_{\mathtt{det}}$, for $k = 0, 1, 2, \ldots$

- Solve *Poisson's eq'n*

$$\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall s \in \mathcal{S}$$

  for (h^{π_k}, C^{π_k}) (system of $|\mathcal{S}|$ eq'ns in $|\mathcal{S}| + 1$ unknowns)
- Check if

$$\max_a \psi^{h^{\pi_k}}(s,a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall s \in \mathcal{S}$$

- If not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s,a)$

Dual LP of OPT properties

Facts:

1. $\exists \;\; \pi^* \in \Pi_{\texttt{det}}$ given by $\pi^*(s) = \arg\max_a \psi^{h^*}(s,a)$, for which $\phi^{\pi*} = \phi^*$
2. (h^*, C^*) are dual-optimal iff $\max_a \psi^{h^*}(s,a) = h^*(s) + C$ for all s.

Policy Iteration: Input initial $\pi_0 \in \Pi_{\texttt{det}}$, for $k = 0, 1, 2, \ldots$

- Solve *Poisson's eq'n*

$$\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$$

  for (h^{π_k}, C^{π_k}) (system of $|\mathcal{S}|$ eq'ns in $|\mathcal{S}| + 1$ unknowns)
- Check if

$$\max_a \psi^{h^{\pi_k}}(s,a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$$

- If not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s,a)$

Next page: Proof that update strictly improves objective.

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s, a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s, a)$

Claim: $v_{\mathtt{avg}}^{\pi_k} < v_{\mathtt{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s, a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s, a)$

Claim: $v_{\mathtt{avg}}^{\pi_k} < v_{\mathtt{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

$v_{\mathtt{avg}}^{\pi_k} = C^{\pi_k}$

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s, a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s, a)$

Claim: $v_{\text{avg}}^{\pi_k} < v_{\text{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

$$v_{\text{avg}}^{\pi_k} = C^{\pi_k} = \sum_s \mu^{\pi_{k+1}}(s) C^{\pi_k}$$

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s,a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s,a)$

Claim: $v_{\text{avg}}^{\pi_k} < v_{\text{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

$$
\begin{aligned}
v_{\text{avg}}^{\pi_k} = C^{\pi_k} &= \sum_s \mu^{\pi_{k+1}}(s) C^{\pi_k} \\
&< \sum_s \mu^{\pi_{k+1}}(s) \left[\psi^{\pi_k}\big(s, \pi_{k+1}(s)\big) - h^{\pi_k}(s) \right]
\end{aligned}
$$

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s,a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall\, s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s,a)$

Claim: $v_{\mathtt{avg}}^{\pi_k} < v_{\mathtt{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

$$\begin{aligned} v_{\mathtt{avg}}^{\pi_k} = C^{\pi_k} &= \sum_s \mu^{\pi_{k+1}}(s) C^{\pi_k} \\ &< \sum_s \mu^{\pi_{k+1}}(s) \left[\psi^{\pi_k}\left(s, \pi_{k+1}(s)\right) - h^{\pi_k}(s) \right] \\ &= \sum_s \mu^{\pi_{k+1}}(s) \left[\mathbf{r}_{s,\pi_{k+1}(s)} + \sum_{s'} \mathbf{P}_{(s,\pi_{k+1}(s)),s'}\, h^{\pi_k}(s') - h^{\pi_k}(s) \right] \end{aligned}$$

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s, a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s, a)$

Claim: $v_{\mathtt{avg}}^{\pi_k} < v_{\mathtt{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

$$
\begin{aligned}
v_{\mathtt{avg}}^{\pi_k} = C^{\pi_k} &= \sum_s \mu^{\pi_{k+1}}(s) C^{\pi_k} \\
&< \sum_s \mu^{\pi_{k+1}}(s) \left[\psi^{\pi_k}\big(s, \pi_{k+1}(s)\big) - h^{\pi_k}(s) \right] \\
&= \sum_s \mu^{\pi_{k+1}}(s) \left[\mathbf{r}_{s,\pi_{k+1}(s)} + \sum_{s'} \mathbf{P}_{(s,\pi_{k+1}(s)),s'} h^{\pi_k}(s') - h^{\pi_k}(s) \right] \\
&= v_{\mathtt{avg}}^{\pi_{k+1}} + \sum_s \left(\sum_{s'} \mu^{\pi_{k+1}}(s) \mathbf{P}_{(s,\pi_{k+1}(s)),s'} h^{\pi_k}(s') - \mu^{\pi_{k+1}}(s) h^{\pi_k}(s) \right)
\end{aligned}
$$

Monotonic Improvement

Policy Iteration: For a deterministic input policy π_k, first

- solve *Poisson's eq'n* $\psi^{h^{\pi_k}}(s, \pi_k(s)) = h^{\pi_k}(s) + C^{\pi_k}$ for (h^{π_k}, C^{π_k})
- check if $\max_a \psi^{h^{\pi_k}}(s, a) = h^{\pi_k}(s) + C^{\pi_k} \quad \forall s \in \mathcal{S}$
- if not, define updated policy $\pi_{k+1}(s) = \arg\max_a \psi^{h^{\pi_k}}(s, a)$

Claim: $v_{\mathrm{avg}}^{\pi_k} < v_{\mathrm{avg}}^{\pi_{k+1}}$, and procedure converges in finite steps.

$$\begin{aligned}
v_{\mathrm{avg}}^{\pi_k} = C^{\pi_k} &= \sum_s \mu^{\pi_{k+1}}(s) C^{\pi_k} \\
&< \sum_s \mu^{\pi_{k+1}}(s) \left[\psi^{\pi_k}\left(s, \pi_{k+1}(s)\right) - h^{\pi_k}(s) \right] \\
&= \sum_s \mu^{\pi_{k+1}}(s) \left[\mathbf{r}_{s, \pi_{k+1}(s)} + \sum_{s'} \mathbf{P}_{(s, \pi_{k+1}(s)), s'} \, h^{\pi_k}(s') - h^{\pi_k}(s) \right] \\
&= v_{\mathrm{avg}}^{\pi_{k+1}} + \sum_s \left(\sum_{s'} \mu^{\pi_{k+1}}(s) \mathbf{P}_{(s, \pi_{k+1}(s)), s'} \, h^{\pi_k}(s') - \mu^{\pi_{k+1}}(s) h^{\pi_k}(s) \right) \\
&= v_{\mathrm{avg}}^{\pi_{k+1}}
\end{aligned}$$

Discussion

Discussion

- Main computational burden is solving Poisson's equations.

Discussion

- Main computational burden is solving Poisson's equations.
- We haven't used reversibility yet.

Discussion

- Main computational burden is solving Poisson's equations.
- We haven't used reversibility yet.
- Reversibility leads to a simplified policy iteration.

The Theorem

Theorem (Anantharam, 2022)

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected.

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$ be as guaranteed.

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$ be as guaranteed. Consider the following iterative procedure.

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$ be as guaranteed. Consider the following iterative procedure.

Let some initial $\pi_0 \in \Pi_{det}$. For $k = 0, 1, 2 \ldots$

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$ be as guaranteed. Consider the following iterative procedure.

Let some initial $\pi_0 \in \Pi_{det}$. For $k = 0, 1, 2 \ldots$

1. *For π_k compute C^{π_k}.*

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$ be as guaranteed. Consider the following iterative procedure.

Let some initial $\pi_0 \in \Pi_{det}$. For $k = 0, 1, 2 \ldots$

1. *For π_k compute C^{π_k}.*
2. *If $\exists\, s$ such that*

$$\frac{\mathbf{r}_{(s,\pi_k(s))} - C^{\pi_k}}{\rho(s, \pi_k(s))} < \arg\max_a \frac{\mathbf{r}_{(s,a)} - C^{\pi_k}}{\rho(s, a)} \tag{1}$$

   Let a_s denote argmax.

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0, 1]$ be as guaranteed. Consider the following iterative procedure.

Let some initial $\pi_0 \in \Pi_{det}$. For $k = 0, 1, 2 \ldots$

1. *For π_k compute C^{π_k}.*
2. *If $\exists\, s$ such that*

$$\frac{\mathbf{r}_{(s,\pi_k(s))} - C^{\pi_k}}{\rho(s, \pi_k(s)))} < \arg\max_{a} \frac{\mathbf{r}_{(s,a)} - C^{\pi_k}}{\rho(s, a)} \tag{1}$$

   Let a_s denote argmax. Set $\pi_{k+1}(s) = a_s$, and $\pi_{k+1}(s') = \pi_k(s')$ for $s' \neq s$. If no such s, terminate.

The Theorem

Theorem (Anantharam, 2022)

Consider a reversible MDP $\mathcal{M}$, whose canonical graph $\mathcal{G}$ is bi-connected. Let $\mathbf{Q}$ and $\rho : \mathcal{S} \times \mathcal{A} \to (0,1]$ be as guaranteed. Consider the following iterative procedure.

Let some initial $\pi_0 \in \Pi_{det}$. For $k = 0, 1, 2 \ldots$

1. *For π_k compute C^{π_k}.*
2. *If $\exists\, s$ such that*

$$\frac{\mathbf{r}_{(s,\pi_k(s))} - C^{\pi_k}}{\rho(s,\pi_k(s)))} < \arg\max_a \frac{\mathbf{r}_{(s,a)} - C^{\pi_k}}{\rho(s,a)} \tag{1}$$

   Let a_s denote argmax. Set $\pi_{k+1}(s) = a_s$, and $\pi_{k+1}(s') = \pi_k(s')$ for $s' \neq s$. If no such s, terminate.

Then $v_{avg}^{\pi_k} < v_{avg}^{\pi_{k+1}}$, and procedure converges in finite steps to an opt. policy.

Proof?

Main Takeaways

Main Takeaways

- Policy gain is an alternative objective function for policy opt.

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility $+$ a little more, could get a better algorithm

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility $+$ a little more, could get a better algorithm
  - No linear system solving!

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility + a little more, could get a better algorithm
  - No linear system solving!
- **Future Directions**:

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility $+$ a little more, could get a better algorithm
  - No linear system solving!
- **Future Directions**:
  - Finite time bounds for ϵ-approximation of opt. policy? Lower bounds without reversibility? Upper bounds with?

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility + a little more, could get a better algorithm
  - No linear system solving!
- **Future Directions**:
  - Finite time bounds for ϵ-approximation of opt. policy? Lower bounds without reversibility? Upper bounds with?
  - RL version—transition dynamics unknown prior, (ϵ, δ) guarantees...

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
    - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility + a little more, could get a better algorithm
    - No linear system solving!
- **Future Directions**:
    - Finite time bounds for ϵ-approximation of opt. policy? Lower bounds without reversibility? Upper bounds with?
    - RL version—transition dynamics unknown prior, (ϵ, δ) guarantees... **Avg. reward RL still an active area of research!**

Main Takeaways

- Policy gain is an alternative objective function for policy opt.
- Optimizing gain for MDPs seems quite tough
  - Dual formulation led to algorithm requiring $\mathcal{O}(|S|)$ linear system solver
- Under reversibility + a little more, could get a better algorithm
  - No linear system solving!
- **Future Directions**:
  - Finite time bounds for ϵ-approximation of opt. policy? Lower bounds without reversibility? Upper bounds with?
  - RL version—transition dynamics unknown prior, (ϵ, δ) guarantees... **Avg. reward RL still an active area of research!**
  - Learnability—identification of irreducible or reversible MDP, learnibility of ρ

Questions

Thanks!

References

- Ross, S. M. "Introduction to Stochastic Dynamic Programming, 1983."
- Cogill, Randy, and Cheng Peng. "Reversible Markov decision processes with an average-reward criterion." SIAM Journal on Control and Optimization 51.1 (2013): 402-418.
- Puterman, Martin L. Markov decision processes: discrete stochastic dynamic programming. John Wiley and Sons, 2014.
- Sutton, Richard S., and Andrew G. Barto. Reinforcement learning: An introduction. MIT press, 2018.
- Anantharam, Venkat. "Reversible Markov decision processes and the Gaussian free field." Systems and Control Letters 169 (2022): 105382.