The value function polytope

- How does the distribution of policies on the polytope effect learning?
- How does gamma change the shape of the polytope?
- How do the dynamics of GPI partition the policy / value spaces?

Distribution of policies

A potentially interesting question to ask about the polytopes is how the policies are distributed over the polytope. To calculate this analytically, we can use the probability chain rule: $p(f(x)) = \mid \det \frac{\partial f(x)}{\partial x} \mid^{-1} p(x)$. Where we set f to be our value functional and $p(x)$ to be a uniform distribution.

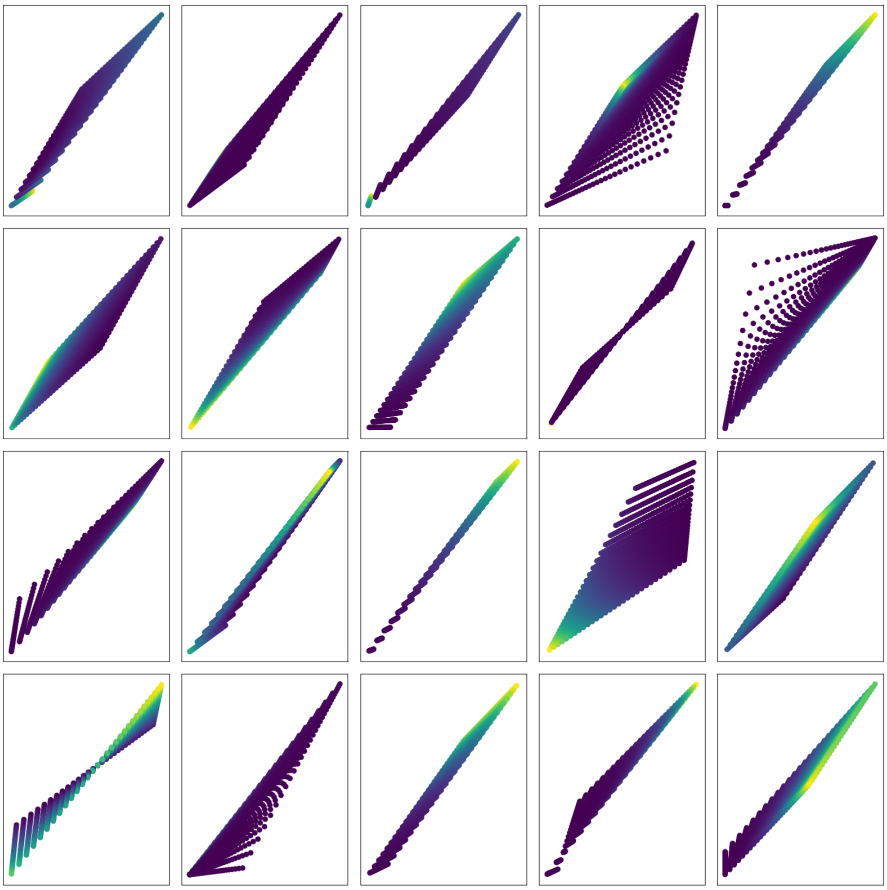

Figure 1: "2-state 2-action MDPs. We have visualised the likelihood of values under a uniform on policies. They are coloured by density. Lighter colour is higher probability"

- **Observation** In some polytopes, many of the policies are close to the optimal policy. In other polytopes, many of the policiesare far away from the optimal policy. **Question** Does this make the MDP harder or easier to solve? **Intuition** If there is a high density near the optimal policy then we could simply sample policies and evaluate them. This would allow us to find a near optimal policy with relative easy.

- **Observation** The density is always concentrated / centered on an edge.
- **Question** how does the entropy of the distribution change under different gamma/transitions/rewards…?

Discounting

How does the shape of the polytope depend on the discount rate? Given an MDP, we can vary the discount rate from 0 to 1 and explore how the shape of the value polytope changes.

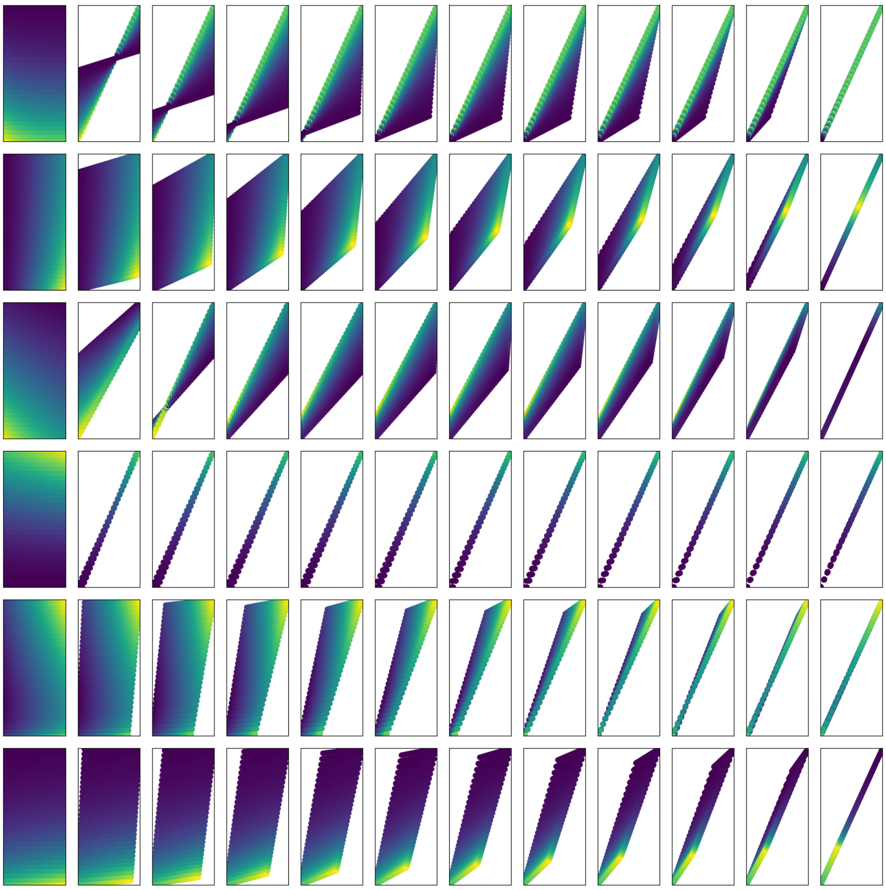

Figure 2: "2-state 2-action MDPs. Here we have shown a few different P/r MDPs and how their polytopes change with changes in discount rate."

- **Observation** As $\gamma \to 1$, all the policies are projected into a 1D space? **Question** Does this make things easier to learn? **Intuition** Orderd 1D spaces are easy to search.

- **Observation** The tranformation that changing the discount applies is quite restricted. They are not generally non-linear, but appear "close to linear", but not quite. **Question** What is the set of functions /transformations that the discount can apply?

Dynamics

(we want to know how much it costs to find the optima)

For each initial policy, we can solve / optimise it to to find the optimal policy (using policy iteration). Here we count how many iterations were required to find the optima (from different starting points / policies).

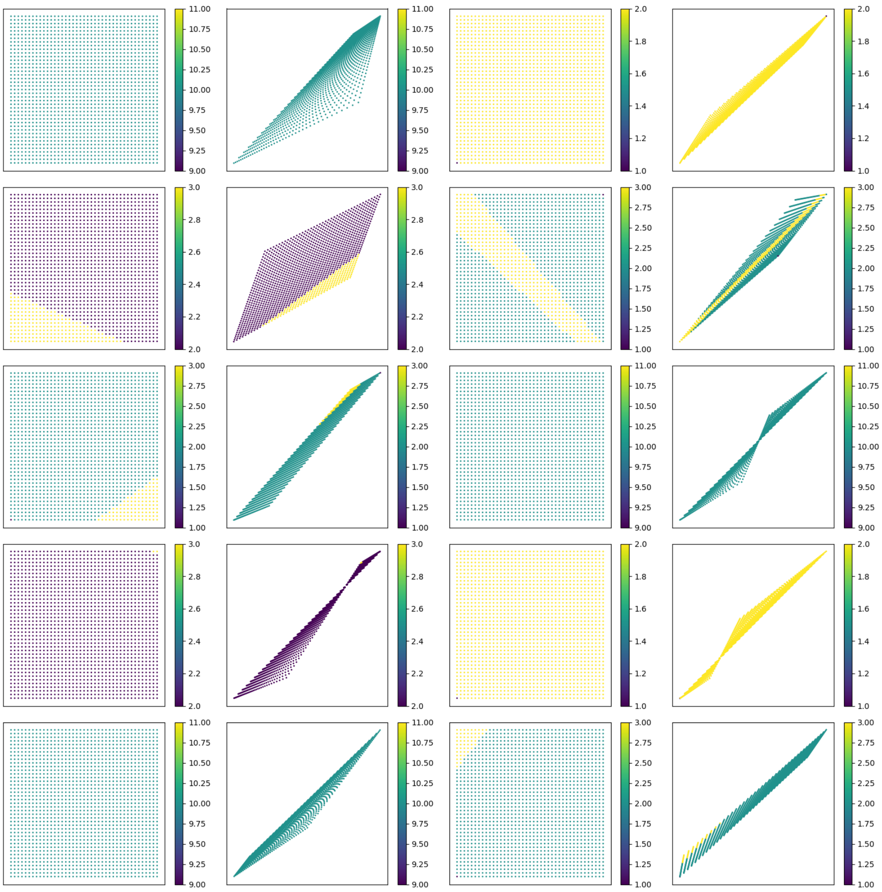

Figure 3: "2-state 2-action MDPs. We have visualised the number of steps required for convergence to the optimal policy. The number of steps are show by color."

- **Observation** Two policies can be within ϵ yet requires more iterations of GPI. **Question** Why are some initial points far harder to solve than others, despite being approximately the same?

- **Observation** With only 2 states and 2 actions, it is possible for 3 partitions to exist. (2,3,4 steps), (2,3,2 steps). **Questions** ???
- **Observation** Sometimes the iterations don't converge. (a bug in the code?)