Each generation

Starting pop Ending pop

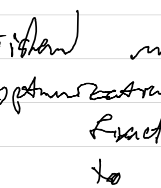
How many different individuals?
How much overlap?
How much functional diversity?
Can we use this as a measure of progression
of learning?
Similar to early stopping

Dry run presentation Aug 29 Tuesday Meeting

Outline paper
outline presentation
30 minutes

meeting time + location email

11 Friday of Reagan
meeting documents
Paper to him by Thursday
morning



Inter-island migrations for reinforcing behavior based interactions
optimization techniques is not necessarily tied to the
functionality of the architecture, but is tied
to its efficacy

Independent behaviors vs blend of behaviors
(multi-objective behavior)
go to ROI avoid collision

vs.

some dynamic blend of these two behaviors

what is actually meant by a "complex interaction"?

mtl
ldca
robot migration

} increase coupling
introduce collision penalties

Agent types

There are two types of agents acting in this environment:

each with a different purpose and dynamics in the environment. Harvesters gather resources, and excavators remove obstacles scattered throughout the environment. Both of these events occur when the appropriate agent type collides with its feature. That is, when a harvester collides with a resource, it gathers the resource, ~~and the global reward is increased by the value of that resource.~~ When an excavator collides with an obstacle, that obstacle is removed from the environment. However, if the wrong type of agent tries to collide with a resource or an obstacle then the action of the agent is discarded, and the agent does not move.

simplifying assumptions

ideal sensors
ideal actuators
homomorphism

— obstacles or resources

Reward Structure

In this work, we are primarily interested in how we can maintain the correct behaviour for agents. However, this may also entail negative reinforcement in the case of a detrimental action. Thus, we explored two reward structures as learning signals for the agents. (These rewards are all global)

The first reward is purely positive reinforcement. Whenever a harvester collects a resource, the global reward

positive rewards for altering the environment in a beneficial way

include penalty when agents collide w/ the wrong type of resource. Excavators do not want to destroy precious resources or harvesters do not want to damage their machinery for colliding with an obstacle debris

Experimental configurations

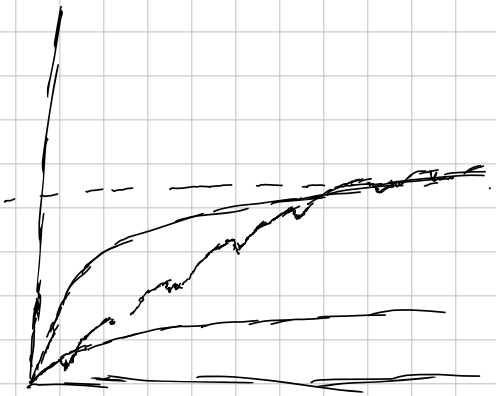
- adjust thresholds
- scale rewards for agents to destroy
- scale rewards for agents to resources
- scale reward penalties
- Transfer learned in one environment to another environment (curriculum)

Allow for resources to spawn dynamically
- Harvesters also learn a poisson distribution for likelihood of resources being present at certain locations
- Bound location of resources to particular areas
- Need another type to scout out these areas

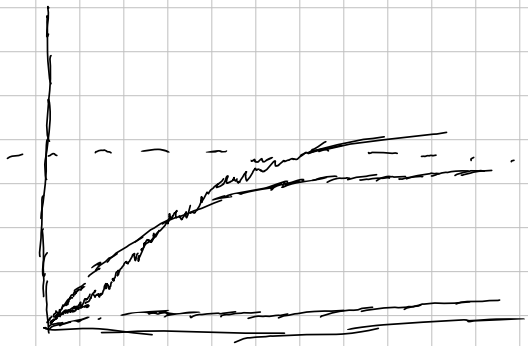
Idea <-> MFI
Ishan made
still issues because...
Intrinsic approach

Proc gen graphics
-> videos of policies acting in env

No collision penalty

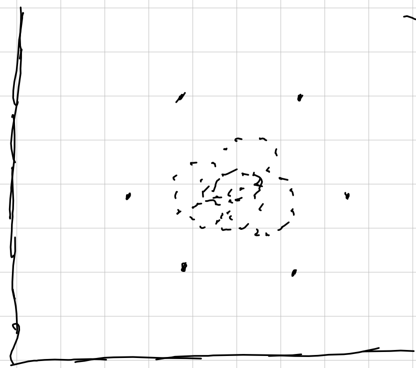


collision penalty



Keep in mind however MFL is pretrained with policies that are explicitly chosen to be beneficial in the environment. The Harvesters are trained to collect resources and the exploiters are trained to remove obstacles, each doing so in the presence of the other type of environment feature.

Show for both harvesters and excavators



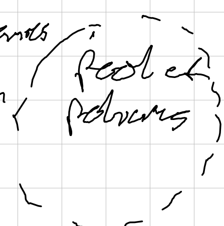
Heatmaps of how common the agent is in that area
- CLEA very limited movement early on
  - less penalty, but less reward
- MPL more movement, but also more collisions
- Infants balance - wait for excavators before traversing bad areas

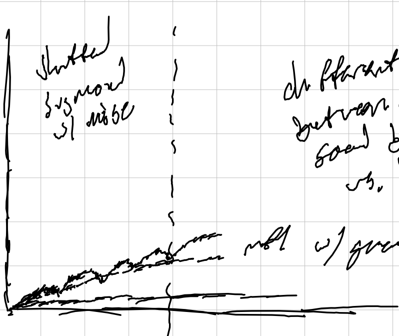
Same as MPL paper for hazardous regions?

- CLEA shuffle order of policy assignment
- might be getting hung up on champions always being in 1, 2, 3 positions
- rand on generation instead of det
- generation into own block
  - check w/ jupyter
  - store generation policy names

importance of time based mitigations
- decaying schedule leads faster overall

Select names rather than policy directly





differentiate between MPL w/ good behaviors vs. bad behaviors

Happens less when there is lower penalty for collisions

Show state views
→ divide into a single view

w/ no penalty, MPL learns on intensity rather than it actually chooses to go into negative regions

Clarity on issue of communication

Slide 12
what is actually being shown?
Loss?

Background slide 19
so what was everything prior?

Slide 33
explain solvent
What do you mean by this?

Slide 47 what is the depth line?

Verify matters credits

Challenges

describe as inherent
-> then soften them
-> get into issues w/
counter examples

Finish your thoughts
before moving on