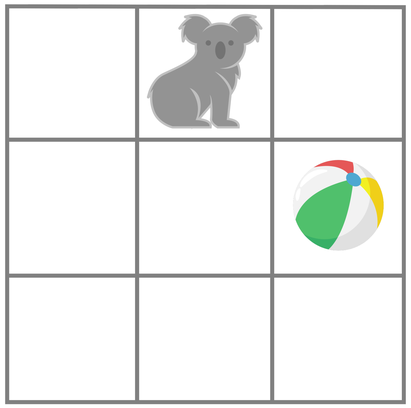
Detecting danger in gridworlds using Gromov's Link Condition

{ Tom Burns and Robert Tang } OIST, Japan & Xi'an Jiaotong–Liverpool University, China

Can we represent a typical AI environment using a single geometric or topological object? And if so, could we recast AI problems as geometric ones?

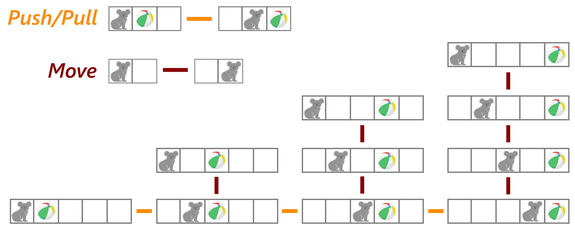
1. Gridworlds



Gridworlds are simplified, grid-like environments in which each *cell* of the grid may be assigned a *label*. The example above shows a 3×3 gridworld with one agent (a koala), one object (a beach ball), and empty floor. Such environments can be used to test and develop AI algorithms, particularly in reinforcement learning [1].

2. State Complexes

A state complex S is a cube complex where vertices (0–cubes) are states of a system, edges (1–cubes) are single reconfigurations of the system, and n–cubes are commuting sets of n independent reconfigurations of the system [2].

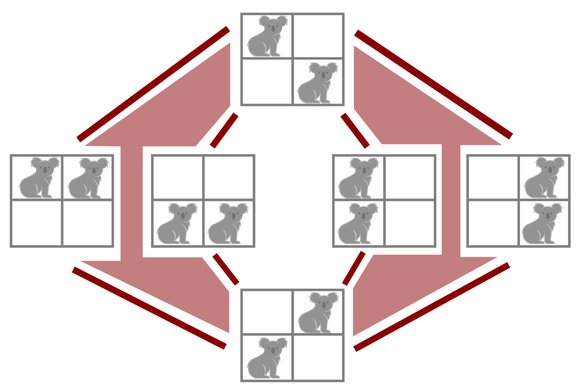


Formally, let the gridworld, G, be a graph. A is a set of possible labels on the vertices of G.

States in S are a choice of labels (chosen from A) for every vertex of G, $s_i : V(G) \to A$.

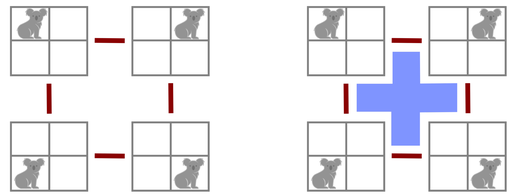
A *generator* ϕ is a collection of three objects:

- the *support*, $SUP(\phi) \subset G$
- the *trace*, $TR(\phi) \subset SUP(\phi)$
- a *relabelling* for the vertex set $TR(\phi)$



3. Dancing with yourself

We modify state complexes to capture agent braiding and more naturally represent the topology. Specially, we fill in 2–cubes whenever we see an agent 'dancing' by itself.

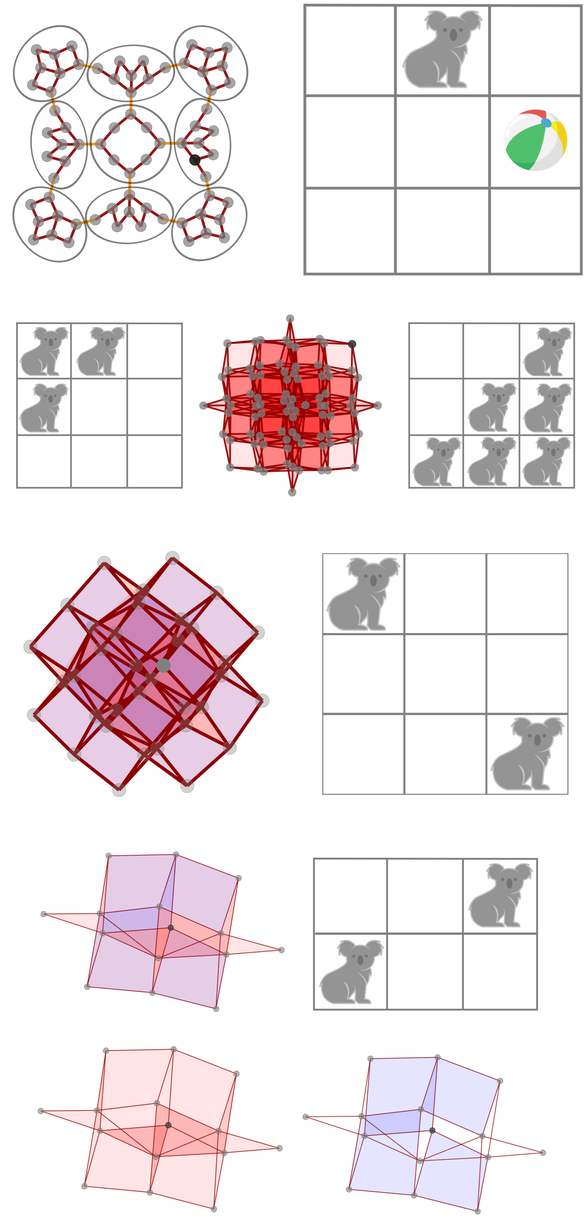


Where they are independent of other generators, these higher-dimensional *dance* generators may commute with other generators. However, doing so can lead to geometric defects (failure of *Gromov's Link Condition*). Serendipitously, we discover these failures occur exactly where undesirable or dangerous states appear in the gridworld.

Theorem. *Let v be a vertex in the modified state complex $\mathcal{S}'$ of an agent-only gridworld. Then*

- *$lk(v)$ satisfies Gromov's Link Condition if and only if it has no empty 2–simplices nor 3–simplices, and*
- *if $lk(v)$ fails Gromov's Link Condition then there exist a pair of agents whose positions differ by either a knight move or a 2–step bishop move.*

4. Small experiments



5. Conclusion

This study presents novel applications of tools from geometric group theory and combinatorics to the AI research community, opening new ways for recasting and analysing AI problems as geometric ones. Using these tools, we show an example of how the intrinsic geometry of a task space serendipitously embeds safety information and makes it possible to determine how far ahead in time an AI system needs to observe to be guaranteed of avoiding dangerous actions.

References

[1] J. Leike et al. AI safety gridworlds. *arXiv*, 1711.09883, 2017.

[2] R. Ghrist and V. Petereson. The Geometry and Topology of Reconfiguration. *Advances in Applied Mathematics*, 38(3):302–323, 2007.

Contact & Paper Information

Paper arXiv:2201.06274
Web tfburns.com
Email thomas.burns@oist.jp
Twitter @tfburns