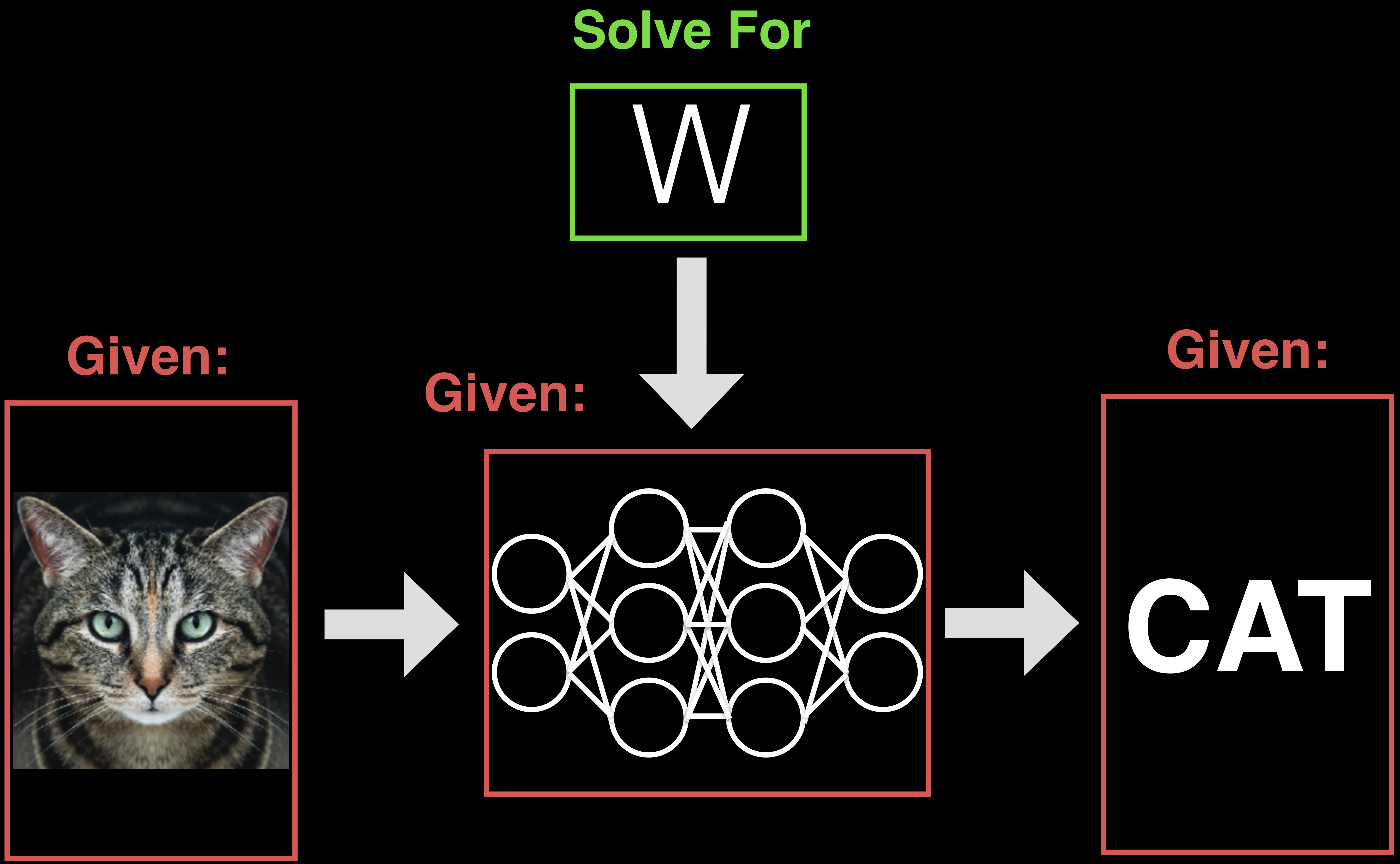
Cryptanalytic Extraction of Neural Network Models

Nicholas Carlini[1], Matthew Jagielski[12], Ilya Mironov[13]

[1]Google, [2]Northeastern, [3]Facebook



Solve For

W

Given:

Given:

Given:

CAT

Our Question:

Given query access to a neural network, can we extract the hidden parameters?

Two views of the problem

Machine Learning
(function approximation)

Mathematical
(direct analysis)

Our Question:

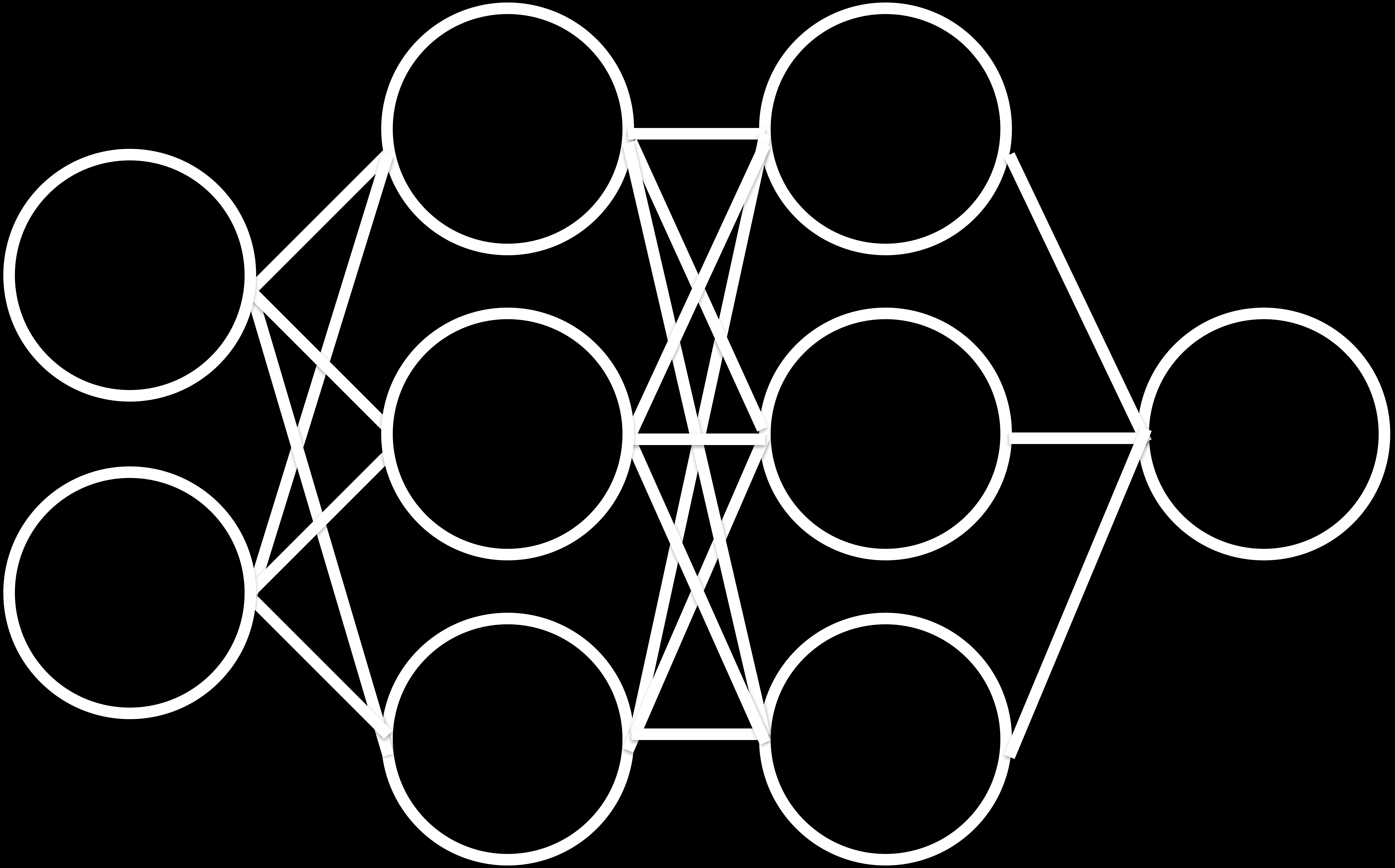
Given query access to a neural network, can we extract the hidden parameters?

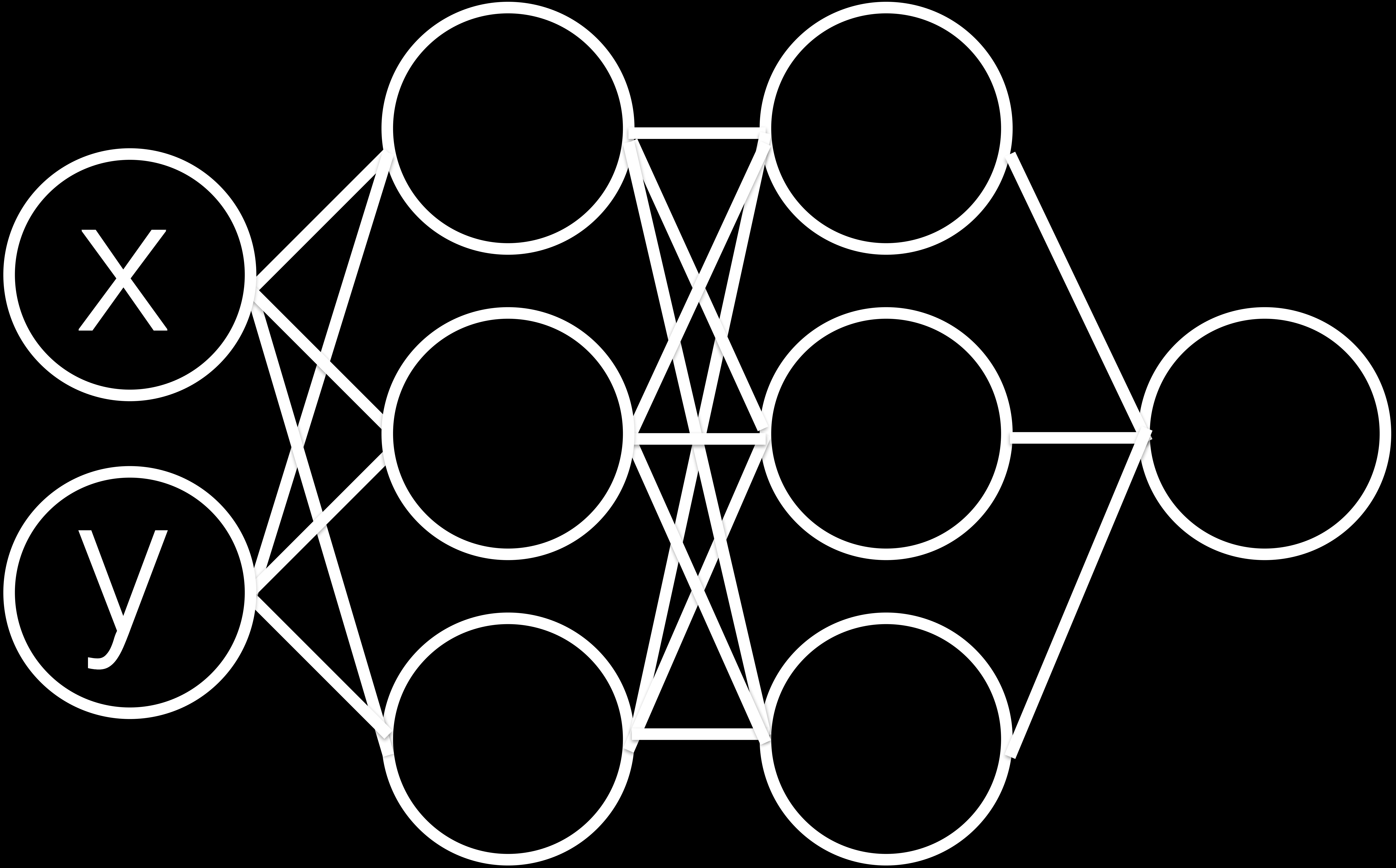
Our Result:

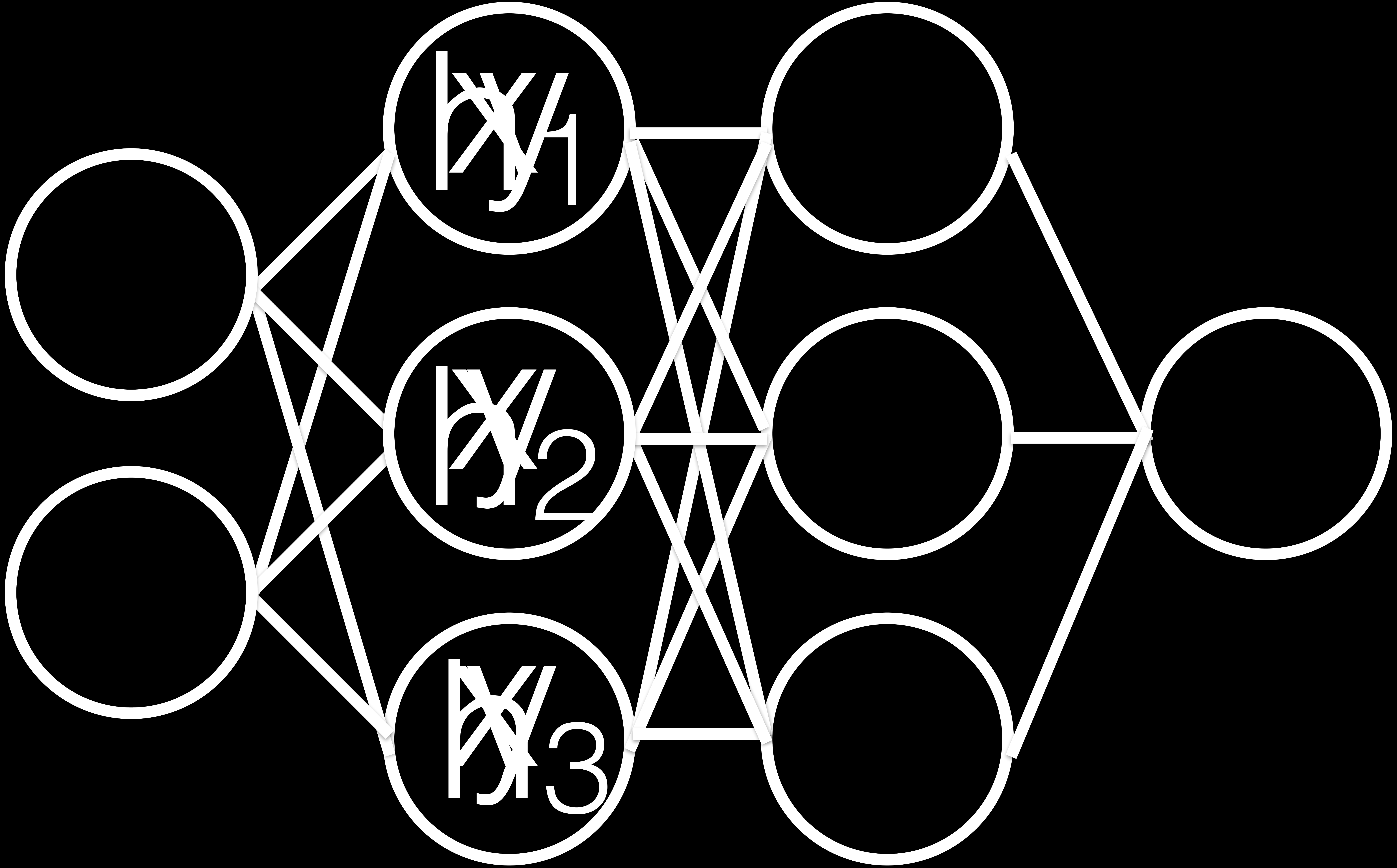
Yes.*

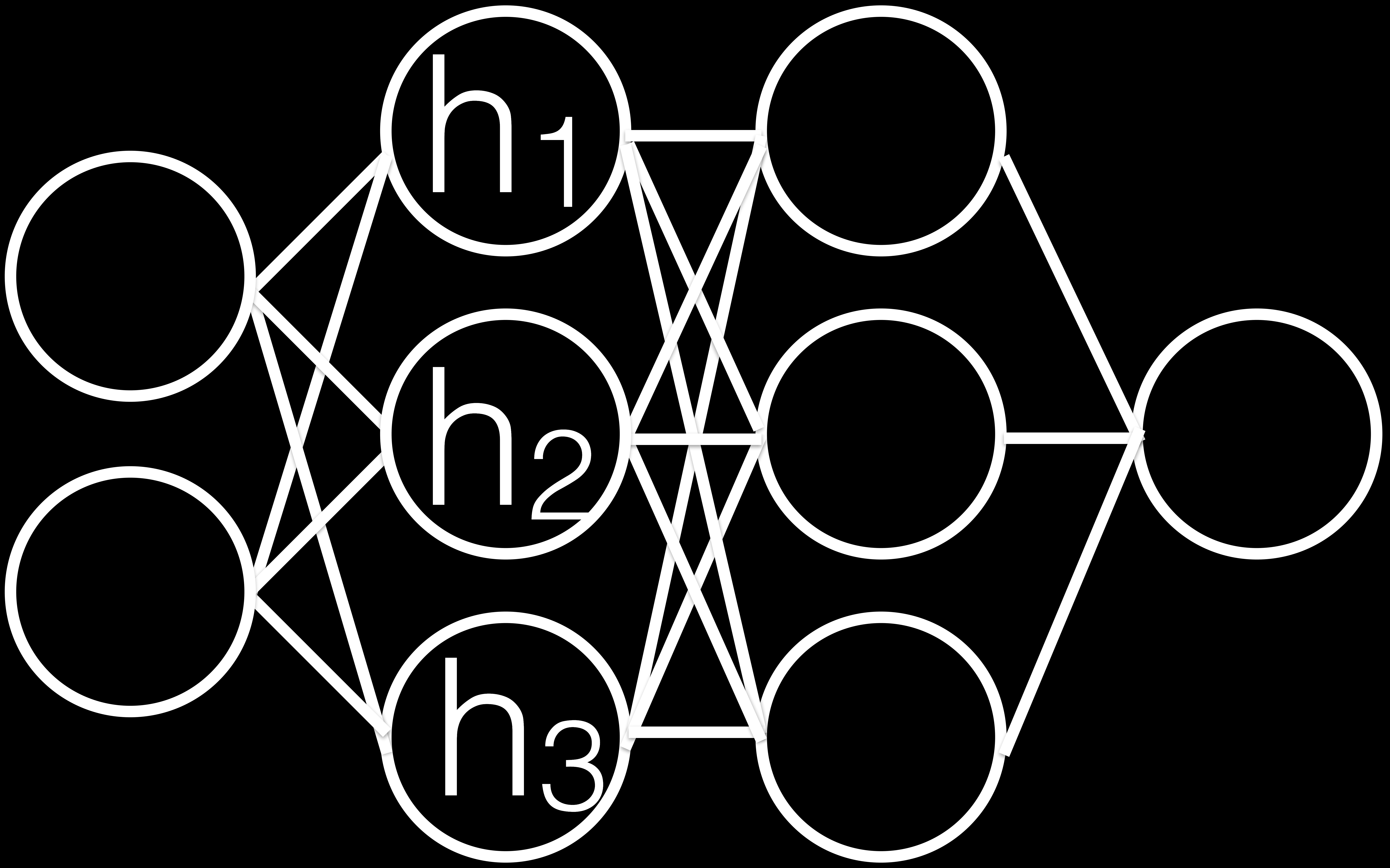
* For small fully connected neural networks with ReLU activations with a few layers evaluated in float64 precision and fully precise inputs and outputs as long as the network isn't pathologically worst-case (e.g., a reduction from 3-SAT) and even then we can only get functional equivalence because exact extraction is provably impossible and even then we only get up to 40 bits of precision when we could theoretically hope for up to 56 bits of precision with float64.

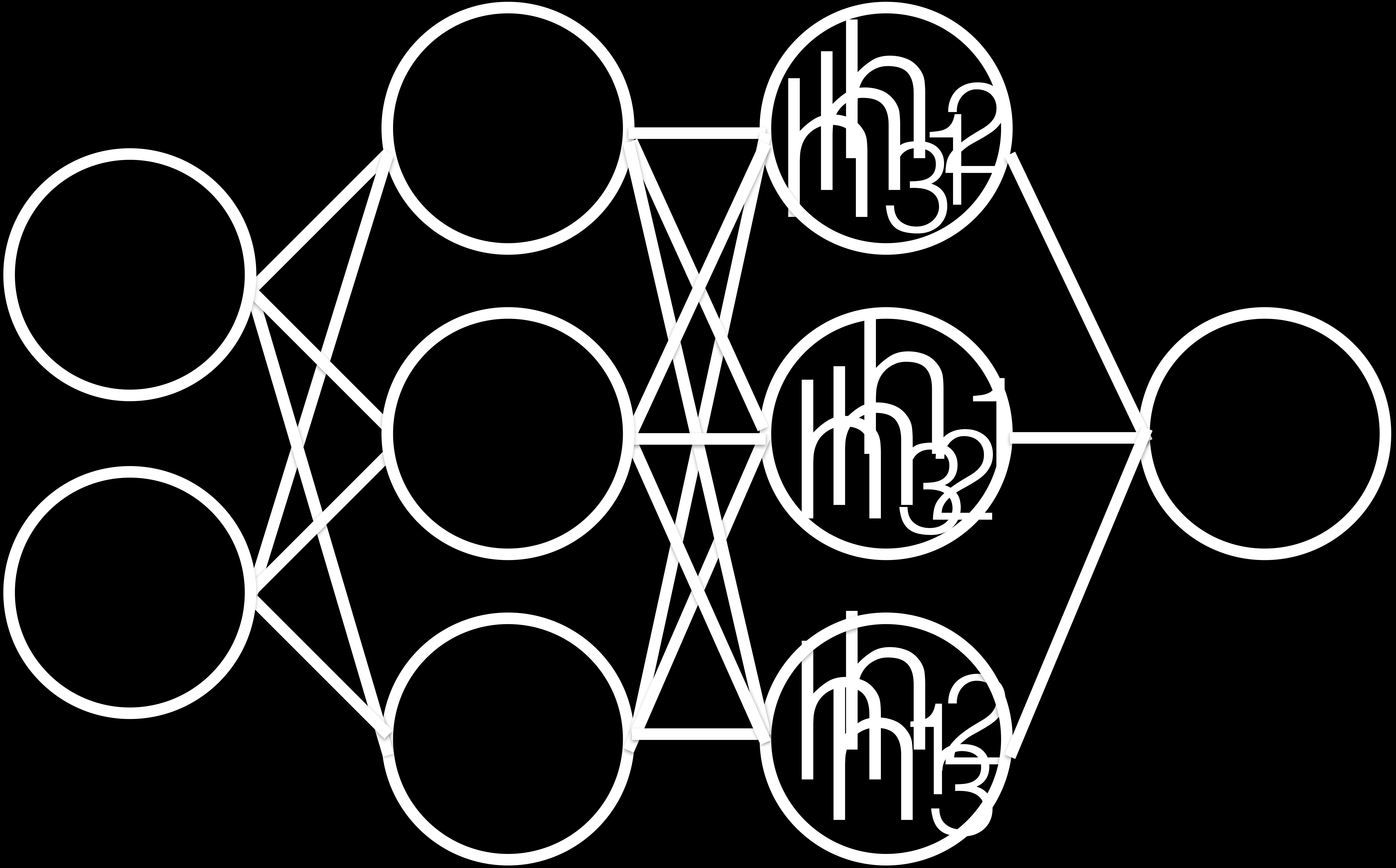
Neural Networks 101

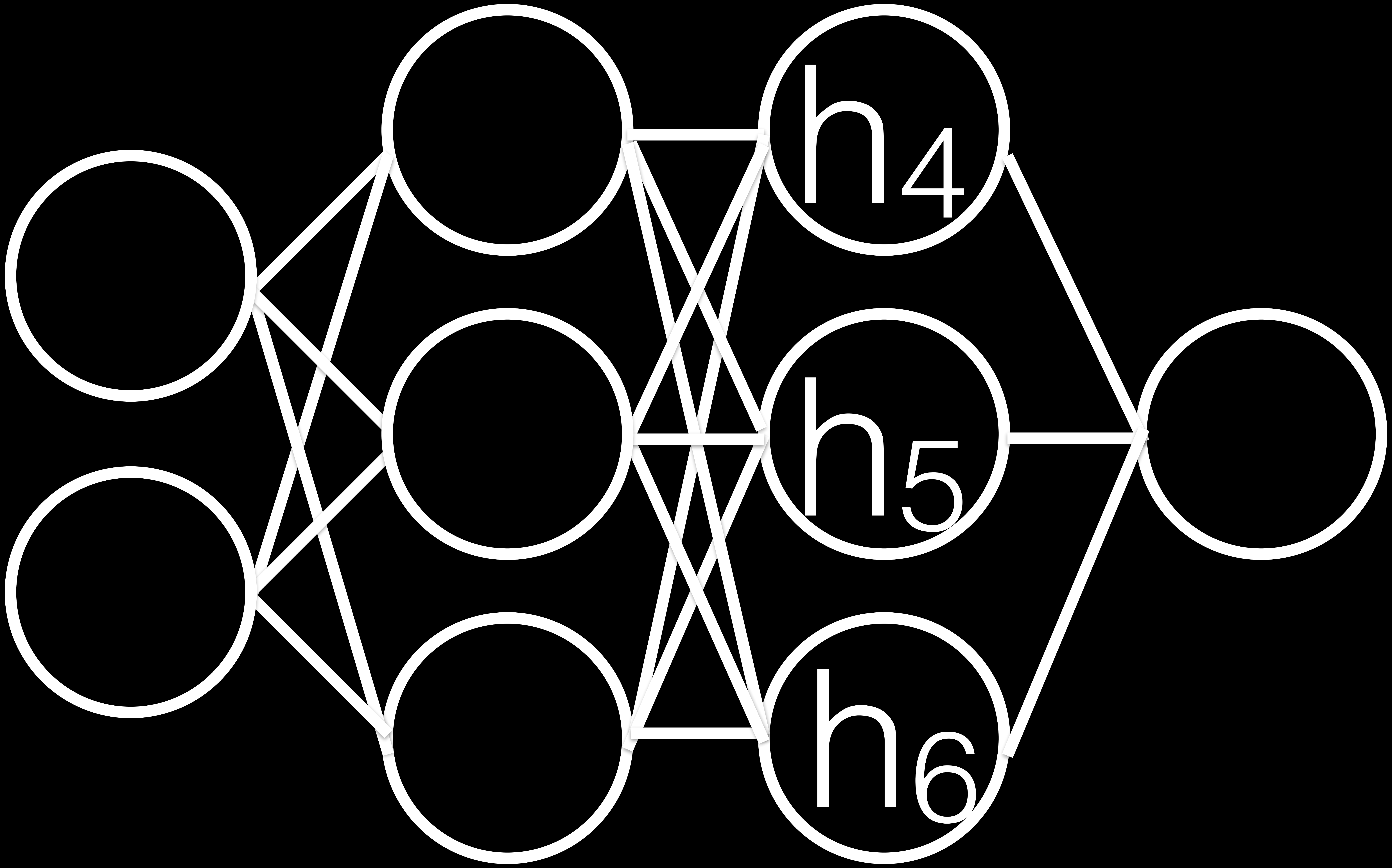


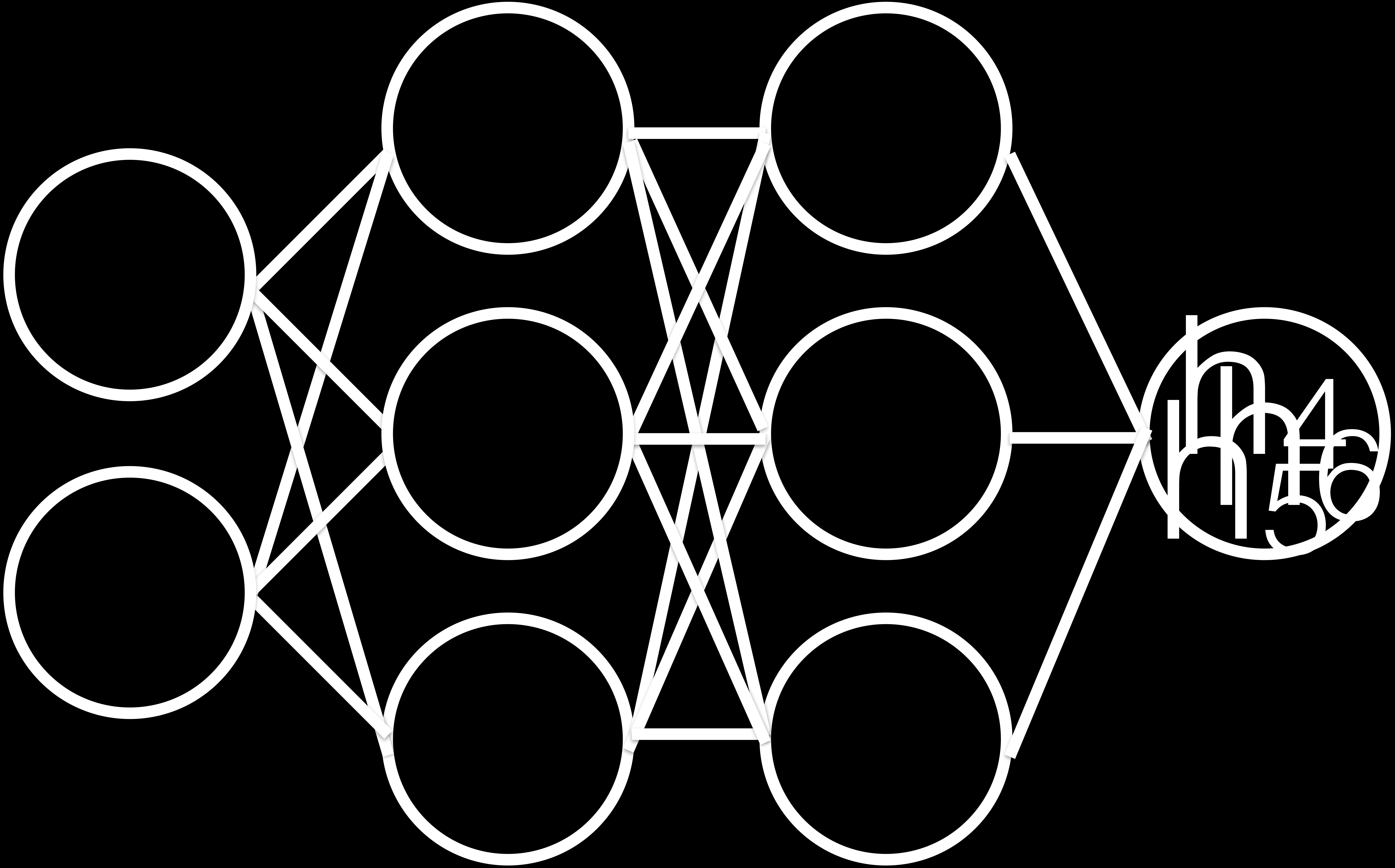

x
y

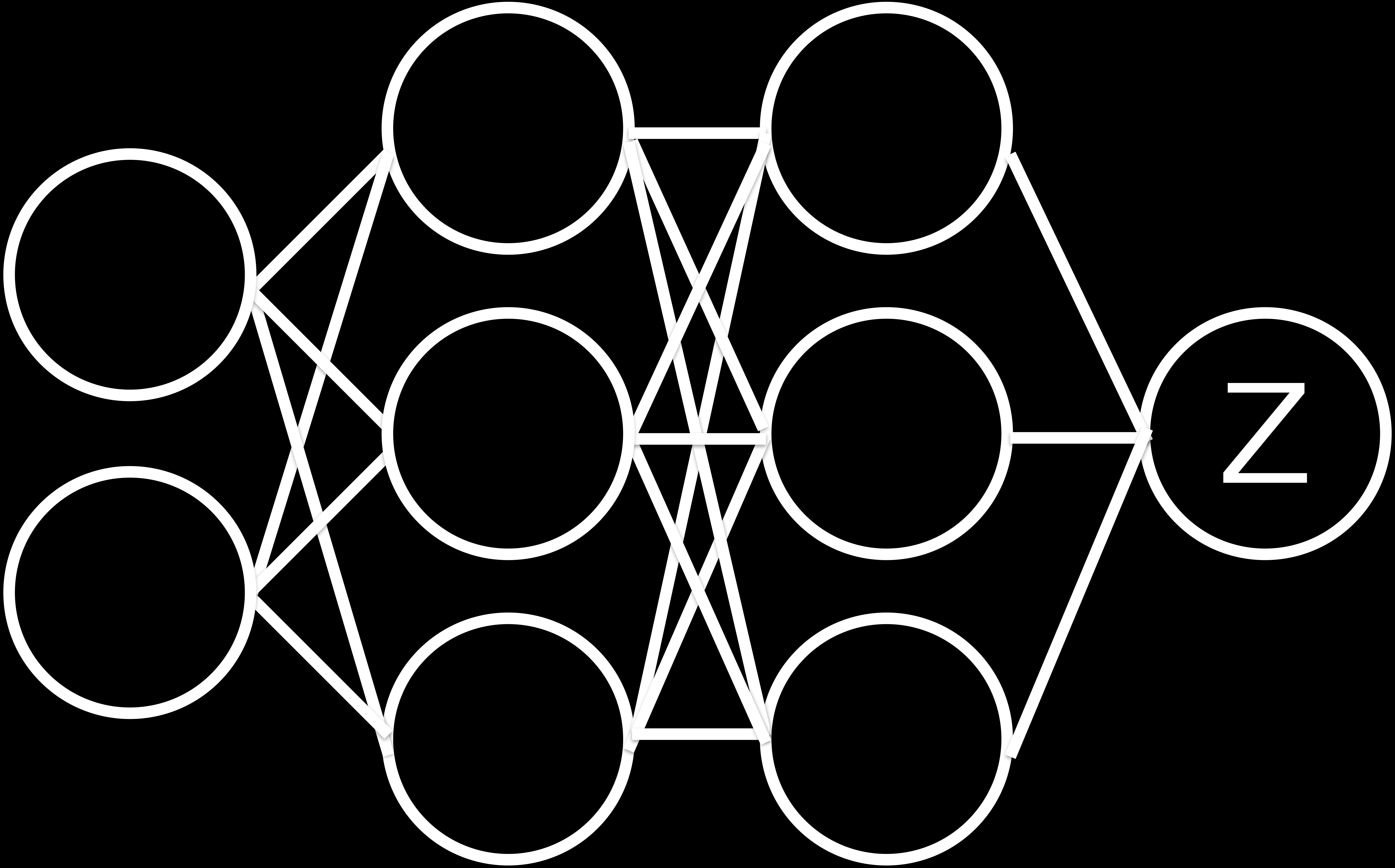

h_1 x_1 y_1
h_2 x_2 y_2
h_3 x_3 y_3

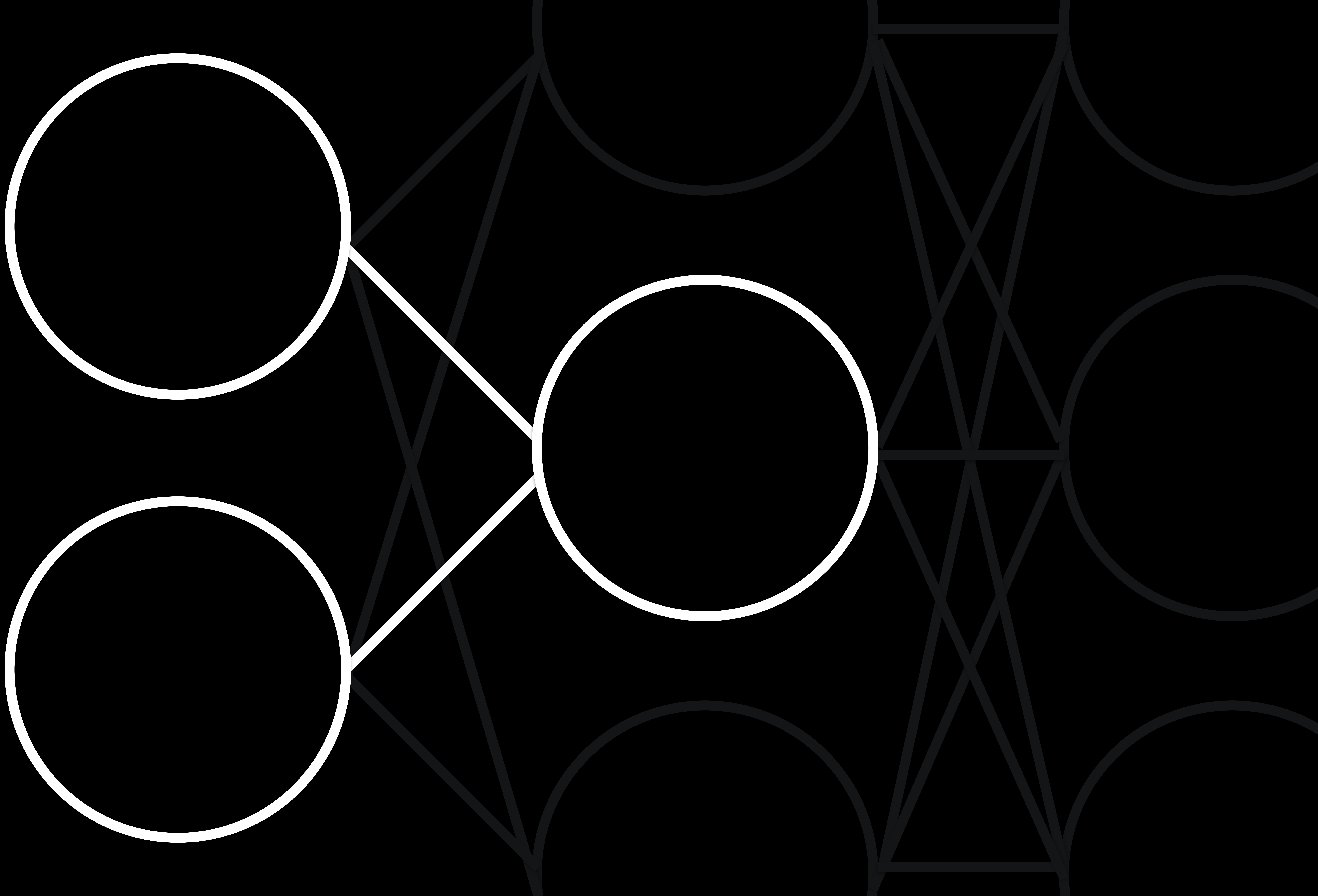

h_1
h_2
h_3

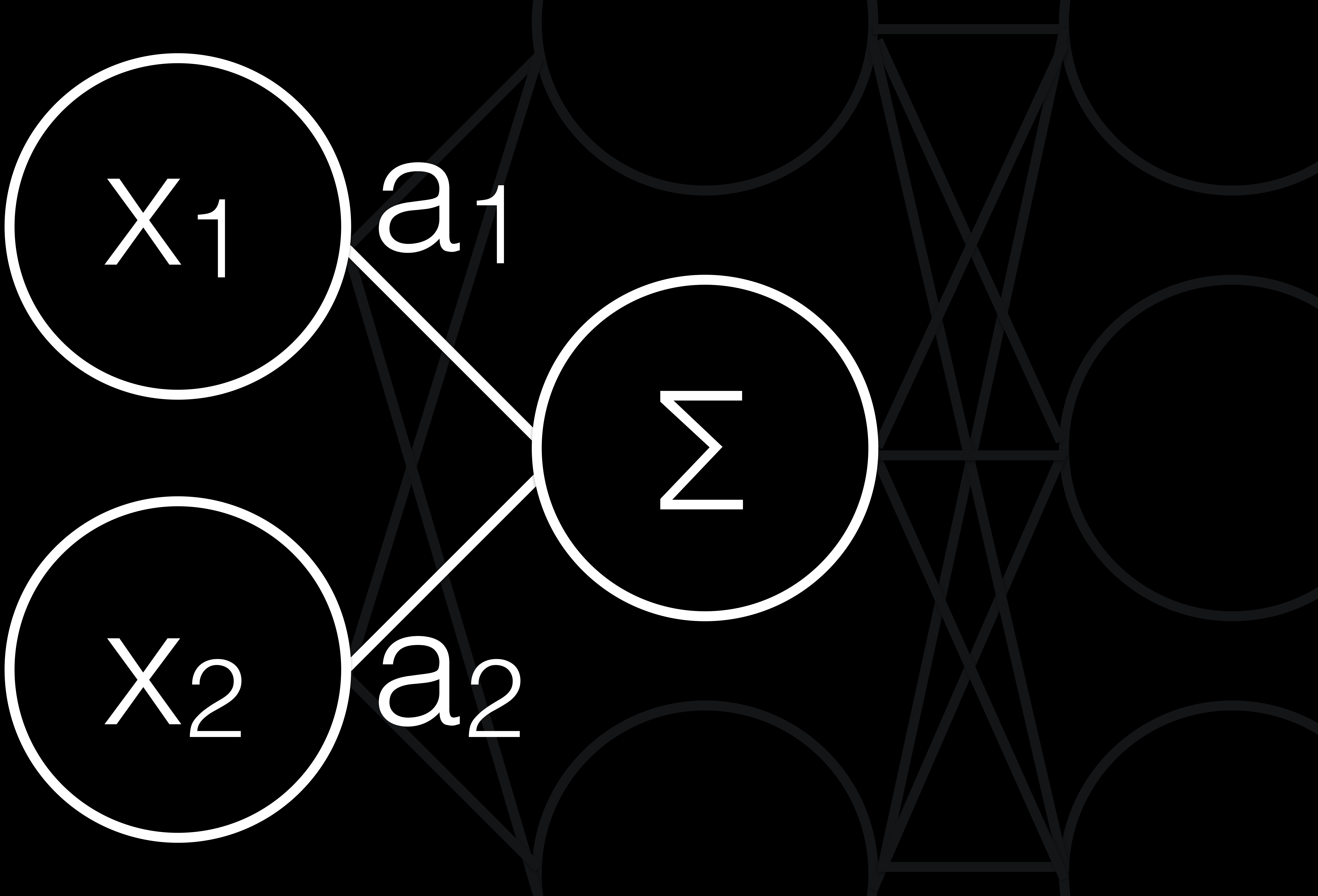

h_2
h_1
h_3
h_1
h_2
h_3
h_2
h_1
h_3

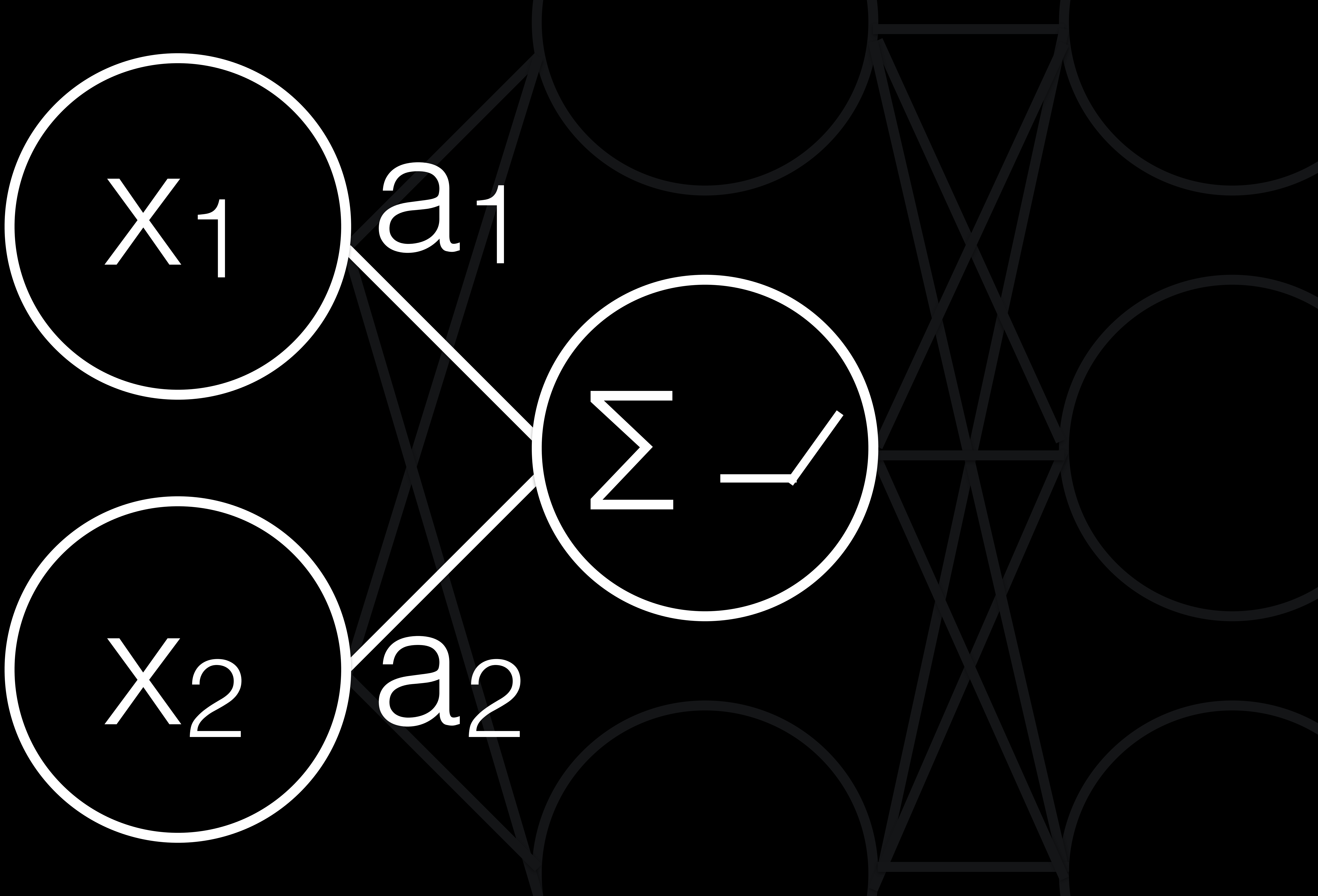

h_4
h_5
h_6

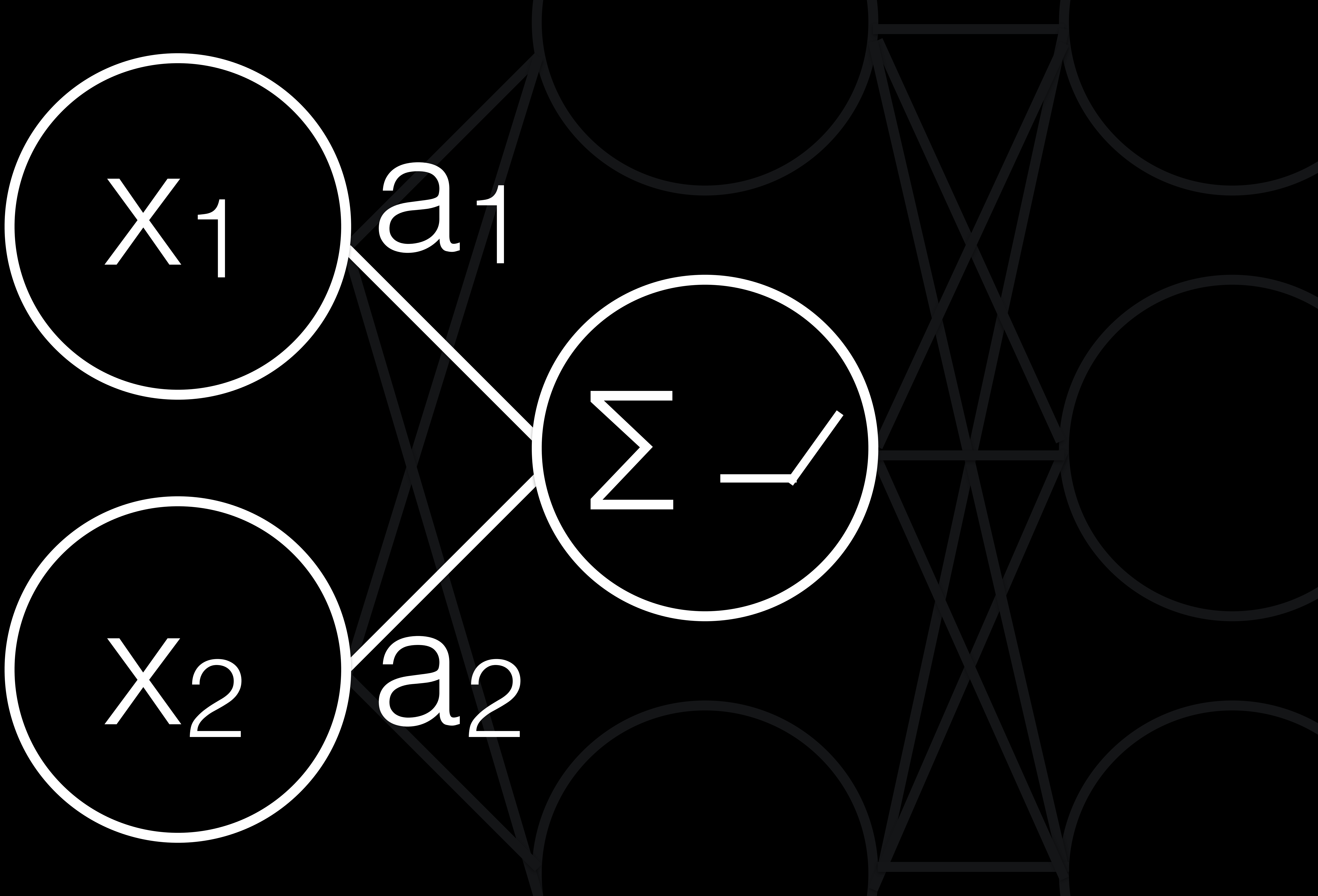

h4
h5
h6


Z



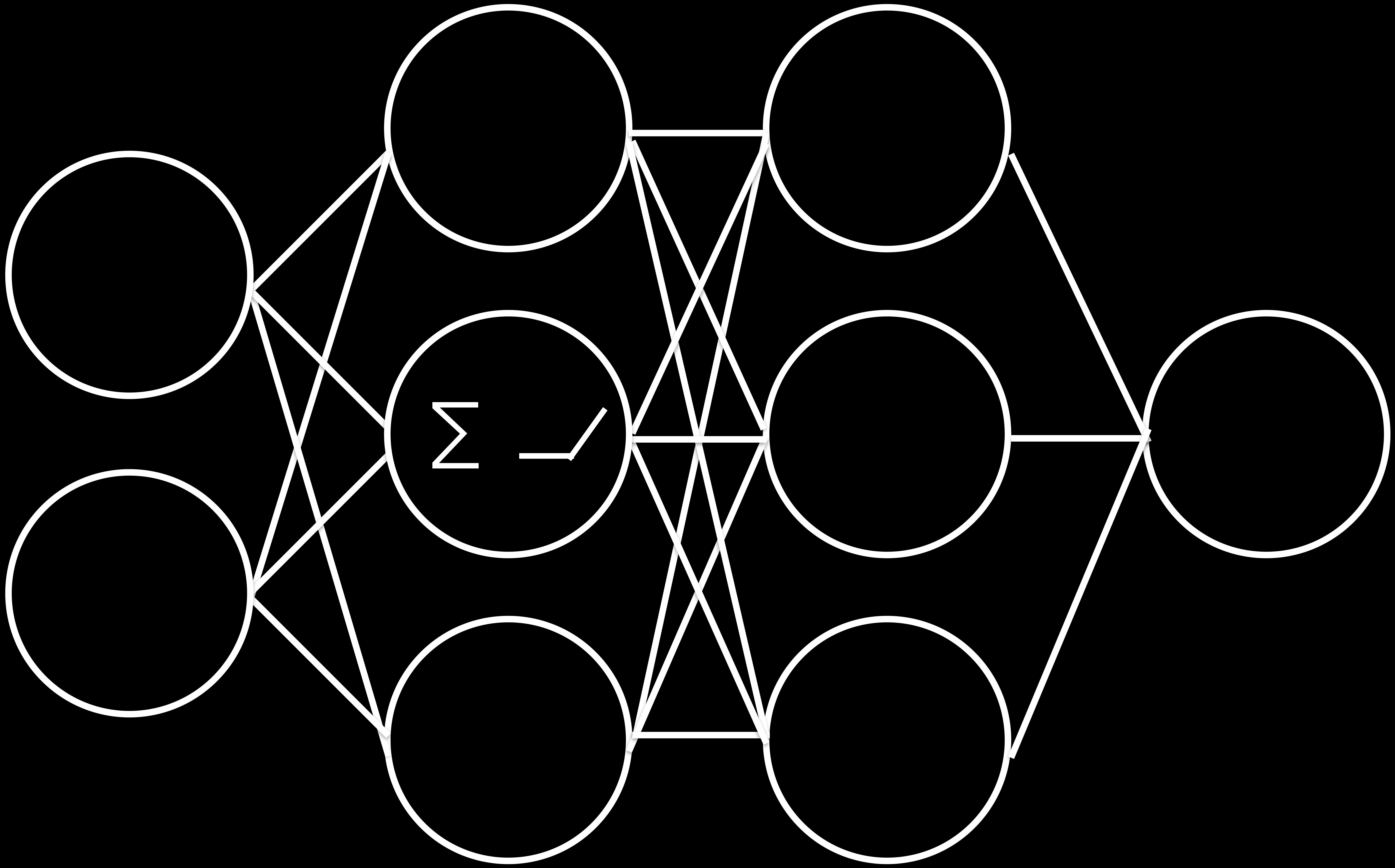

x1
a1
Σ
x2
a2

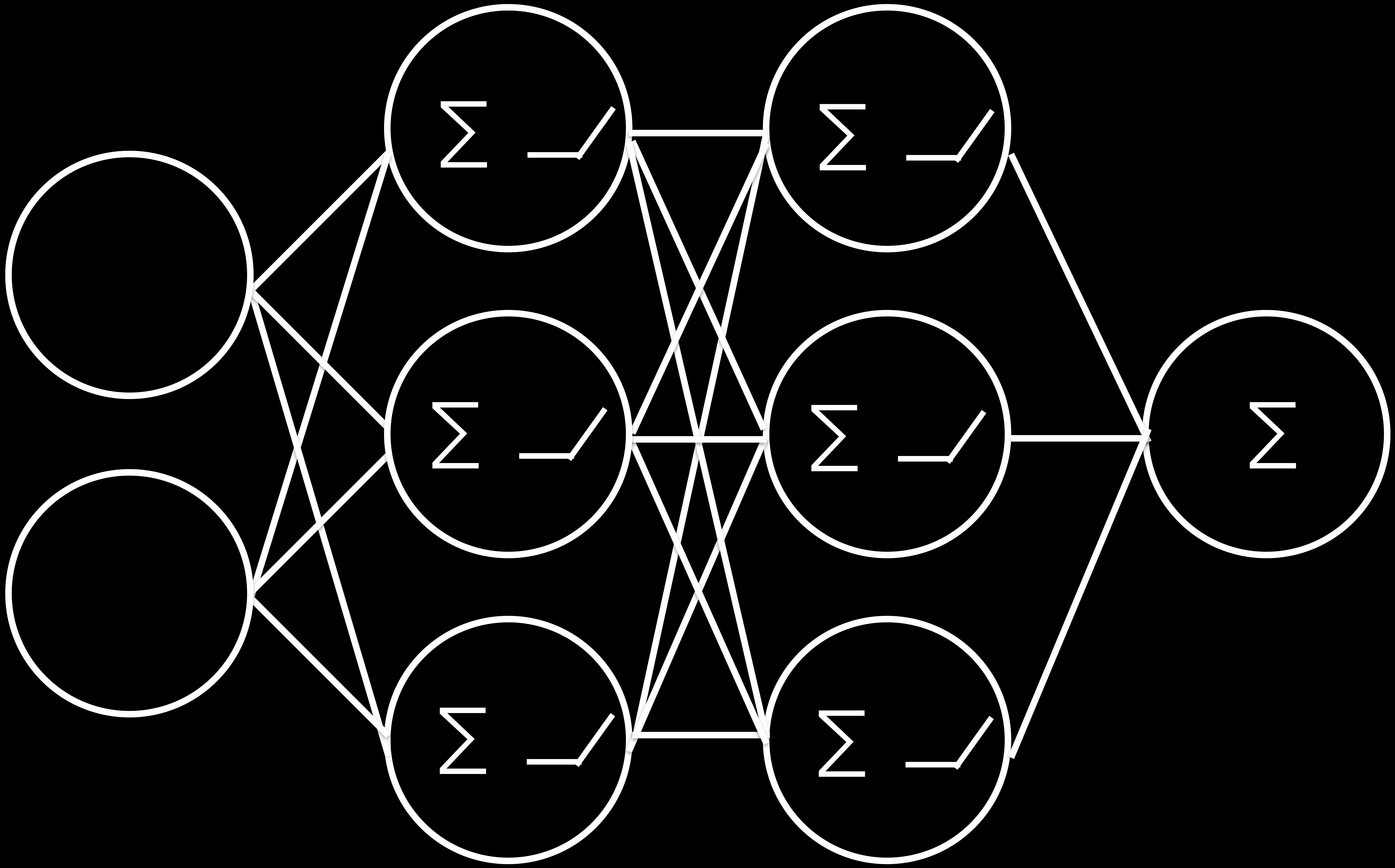

x1
a1
Σ
x2
a2


x1
a1
x2
a2
Σ

$$\mathrm{ReLU}(x) = \max(x, 0)$$

x_1
a_1
Σ
x_2
a_2


Σ


Σ
Σ
Σ
Σ
Σ
Σ
Σ

Extracting Neural Networks

Given (oracle) query access to a neural network, can we extract the *exact* model?

Given (oracle) query access to a neural network, can we extract a *functionally equivalent* model?

Given (oracle) query access to a neural network, can we extract a *functionally equivalent* model?

Given (oracle) query access to a neural network, learned through stochastic gradient descent, can we extract a *functionally equivalent* model?

Given (oracle) query access to a neural network,
learned through stochastic gradient descent,
can we extract a *functionally equivalent* model?

This paper: **yes** (empirically)

[MSDH19, JCB+20]

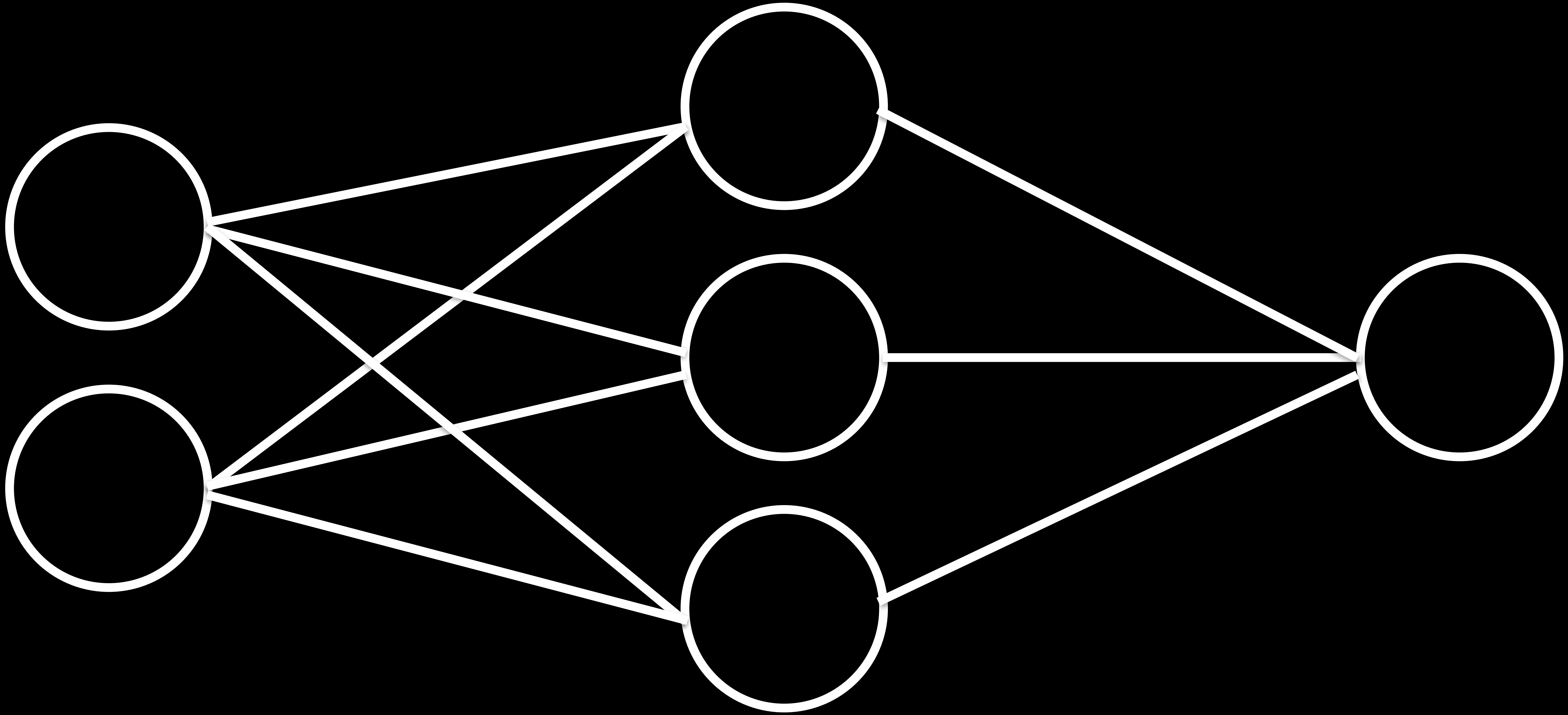
Reduced Round Attack:

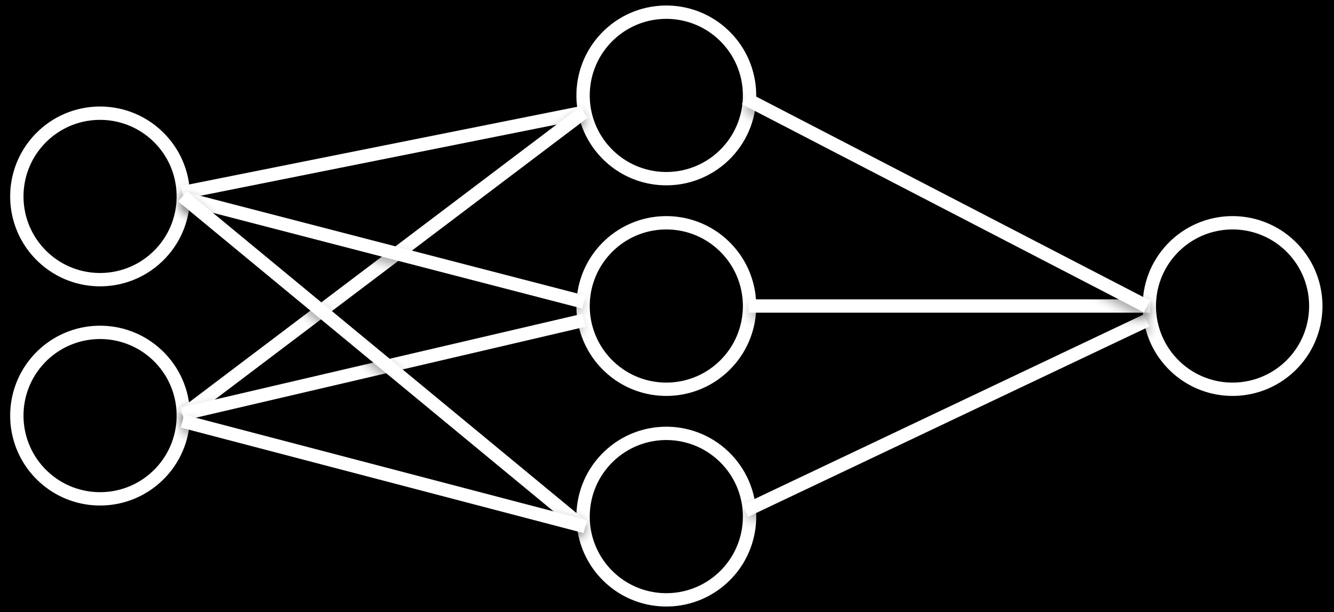
1 Hidden Layer

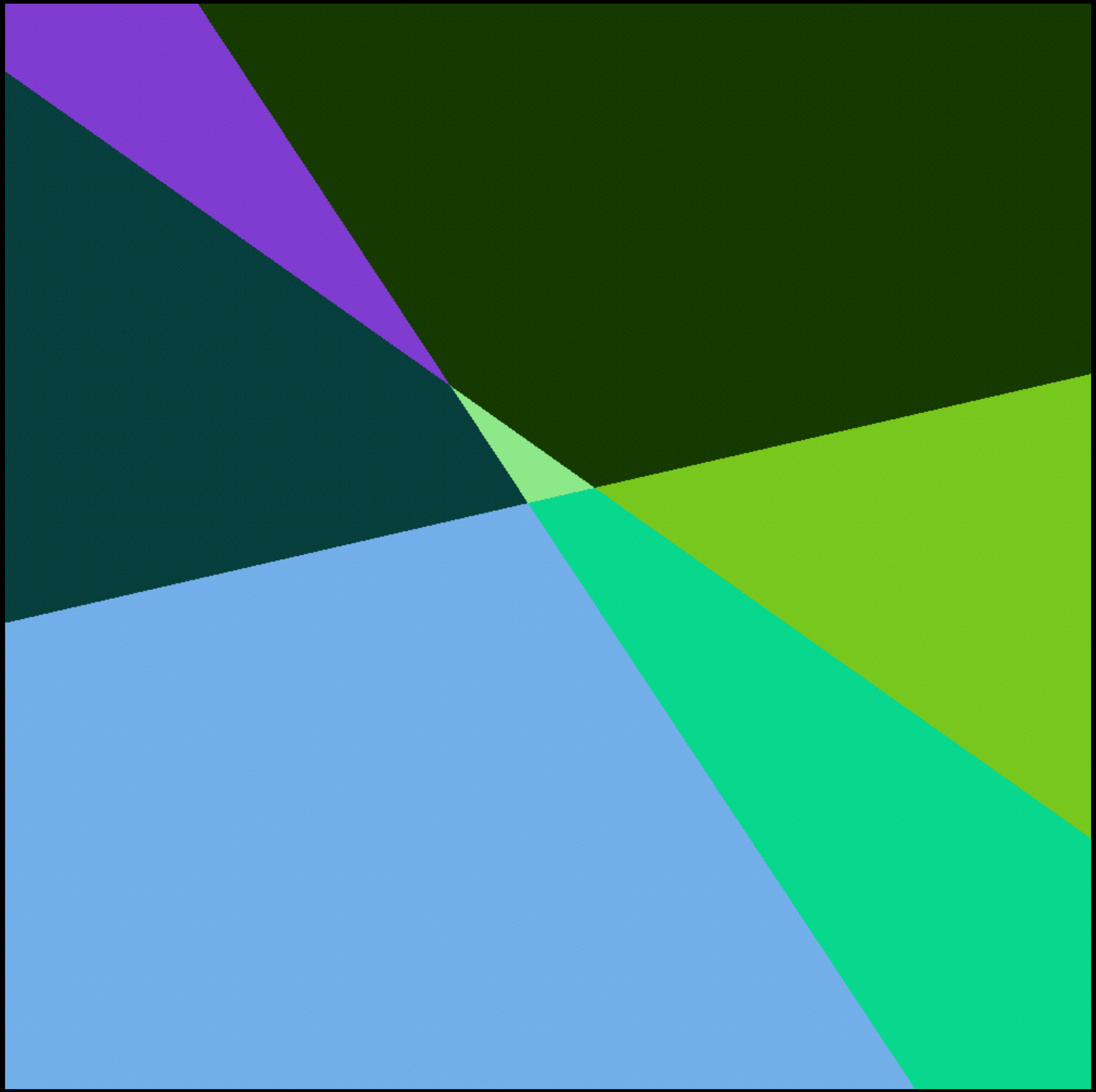
[MSDH19, JCB+20]

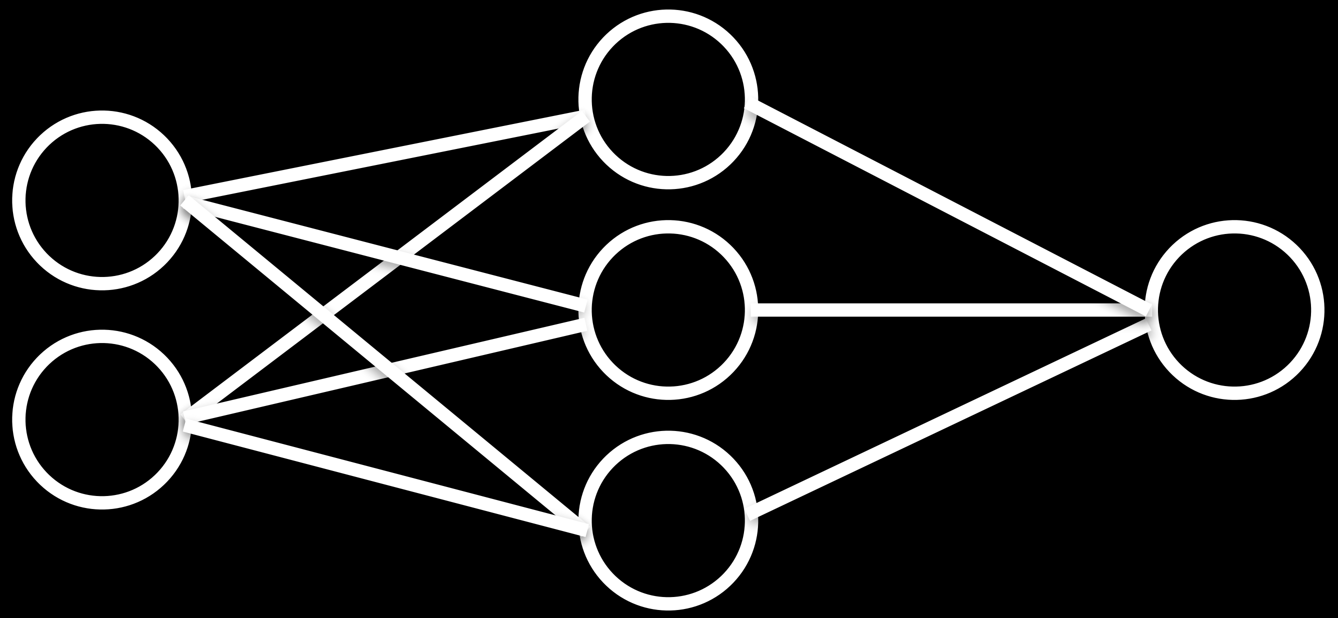
Visual Intuition

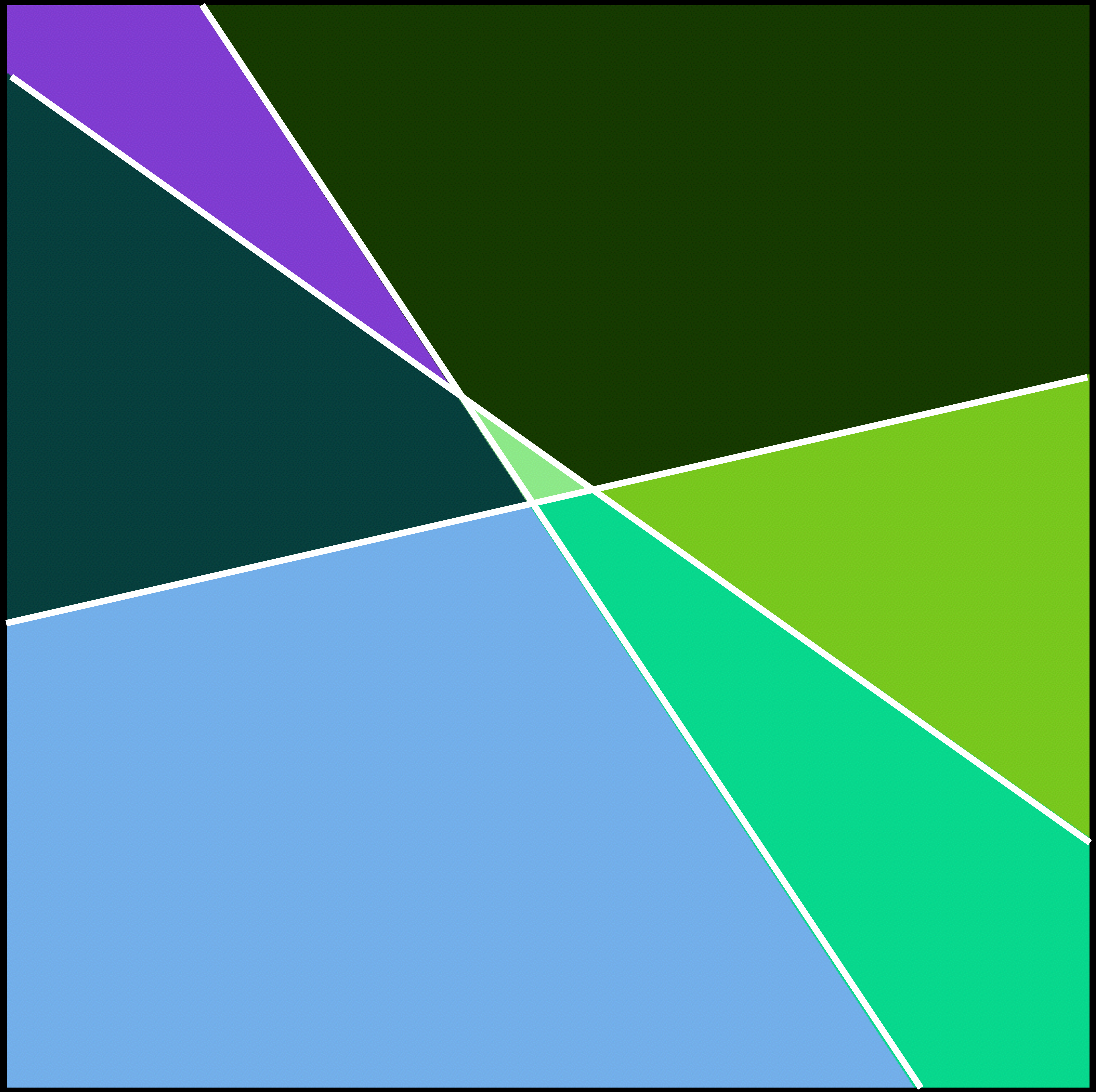
[MSDH19, JCB+20]

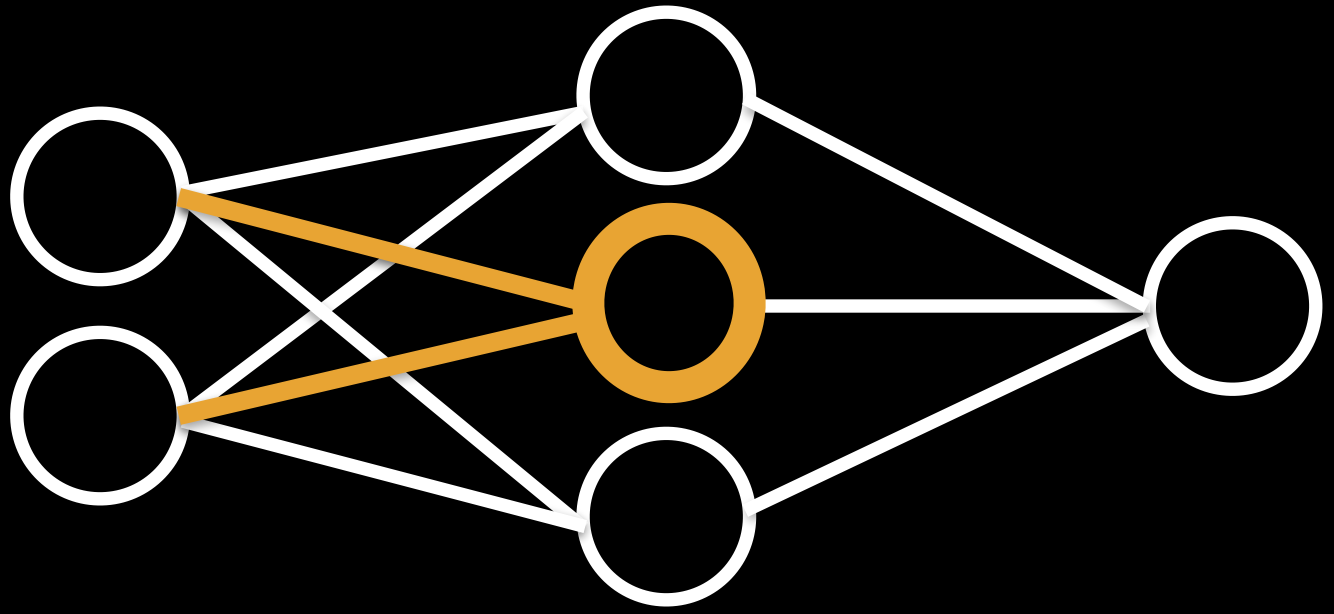


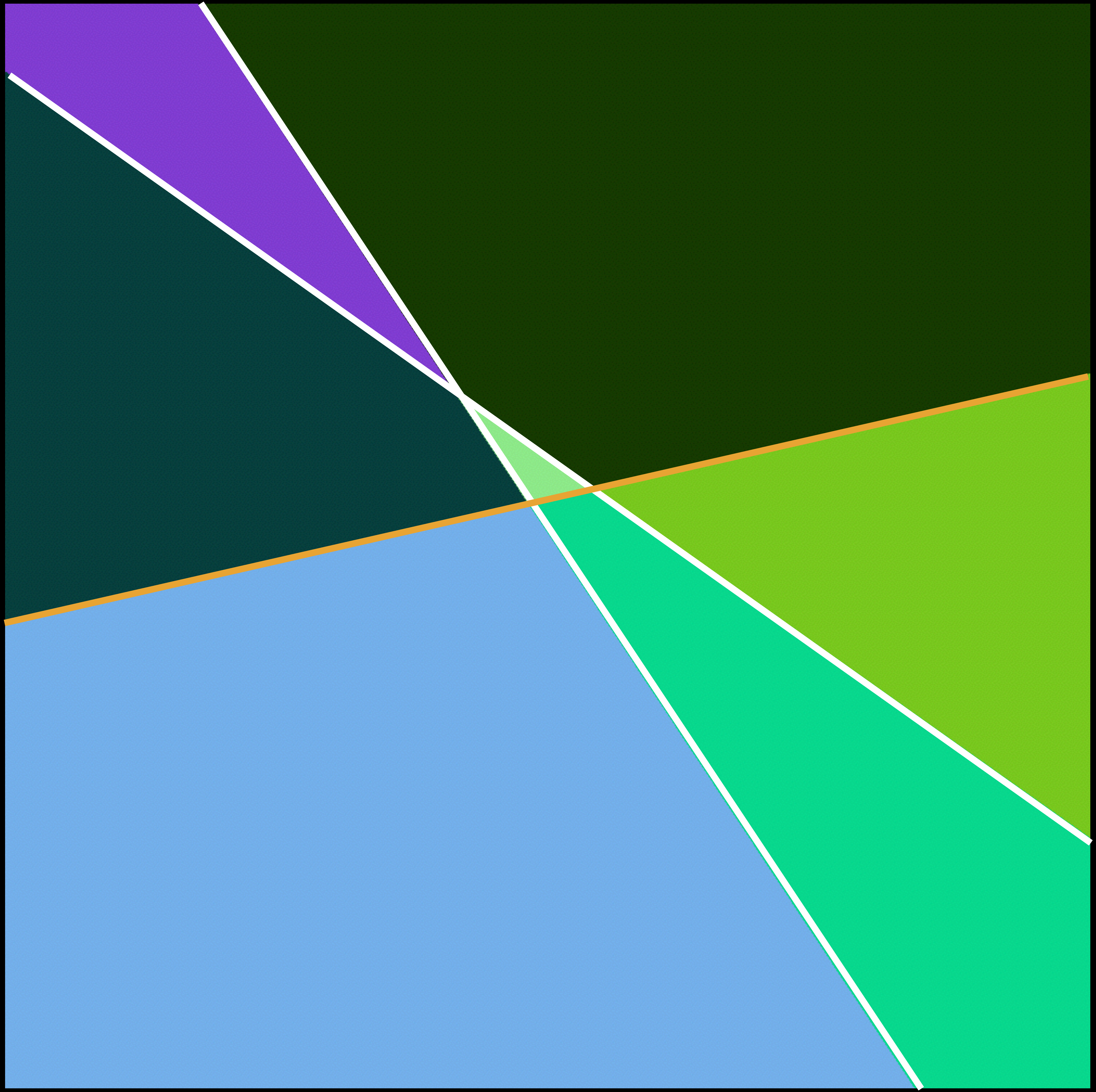


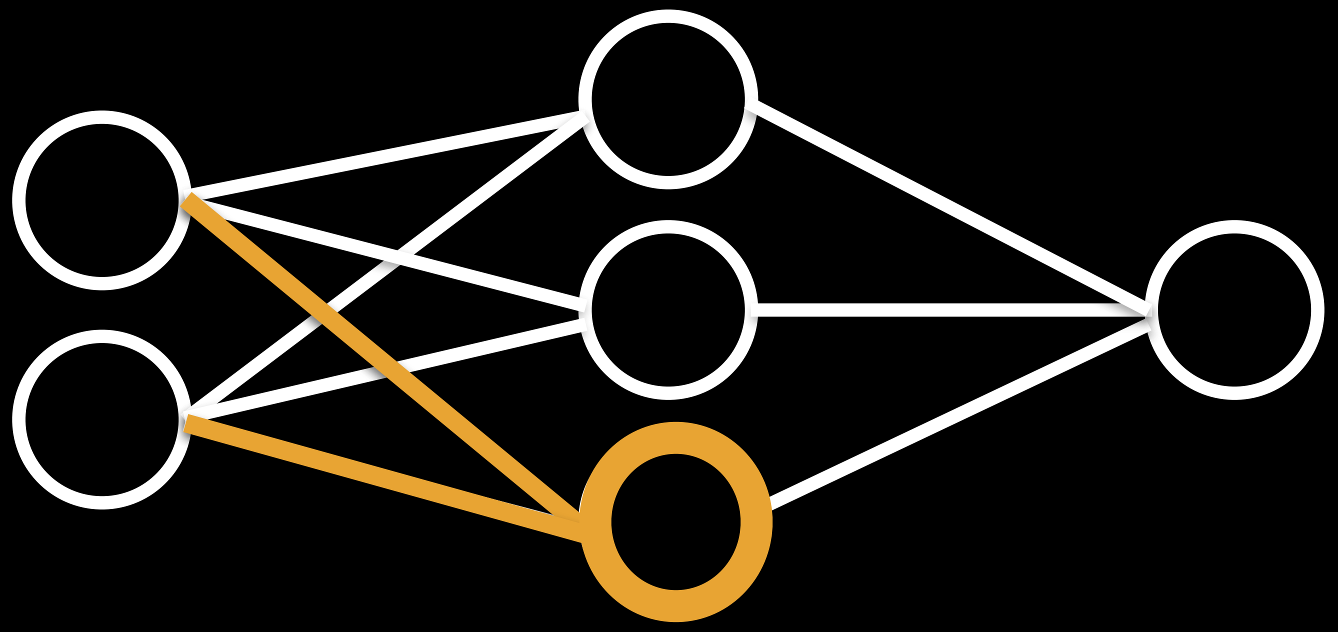


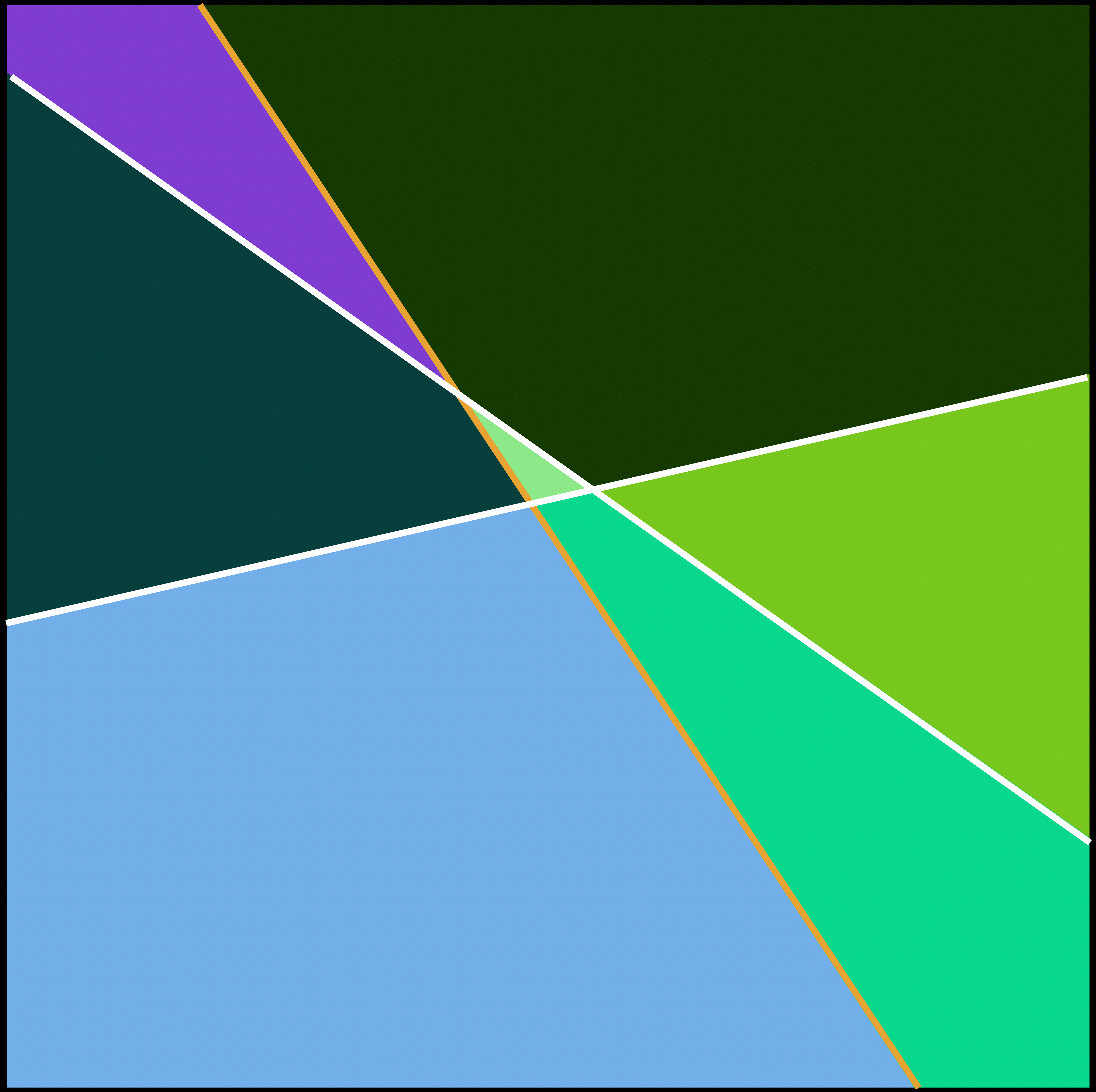


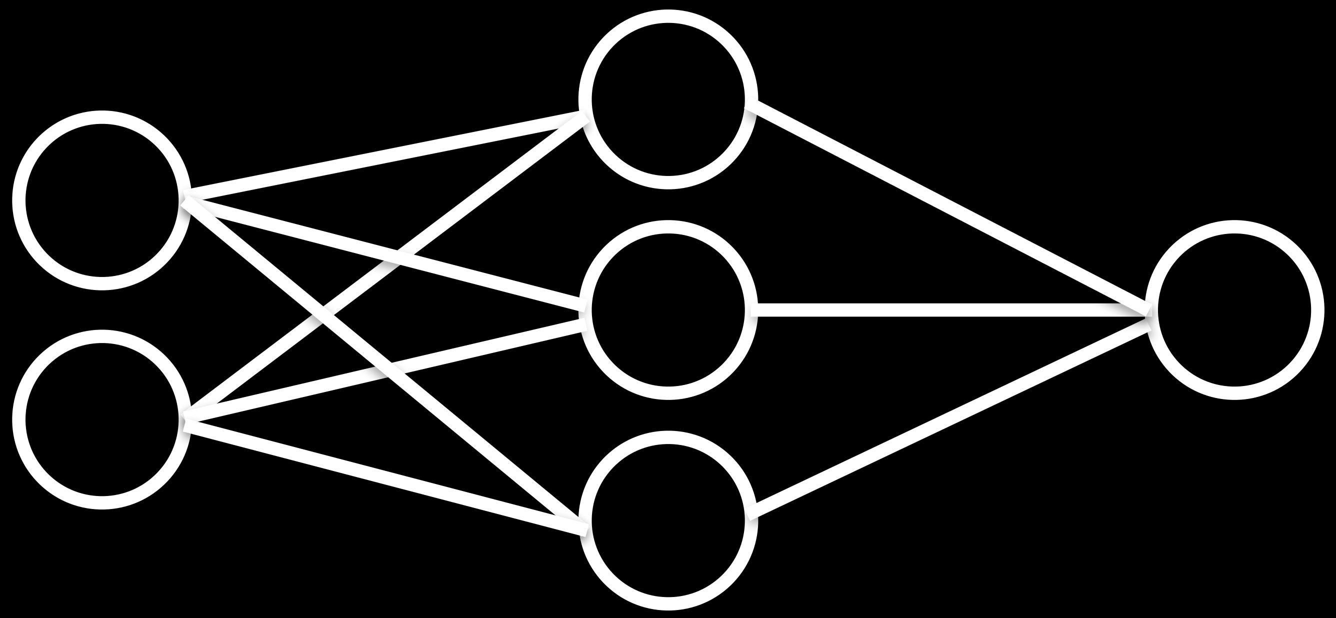


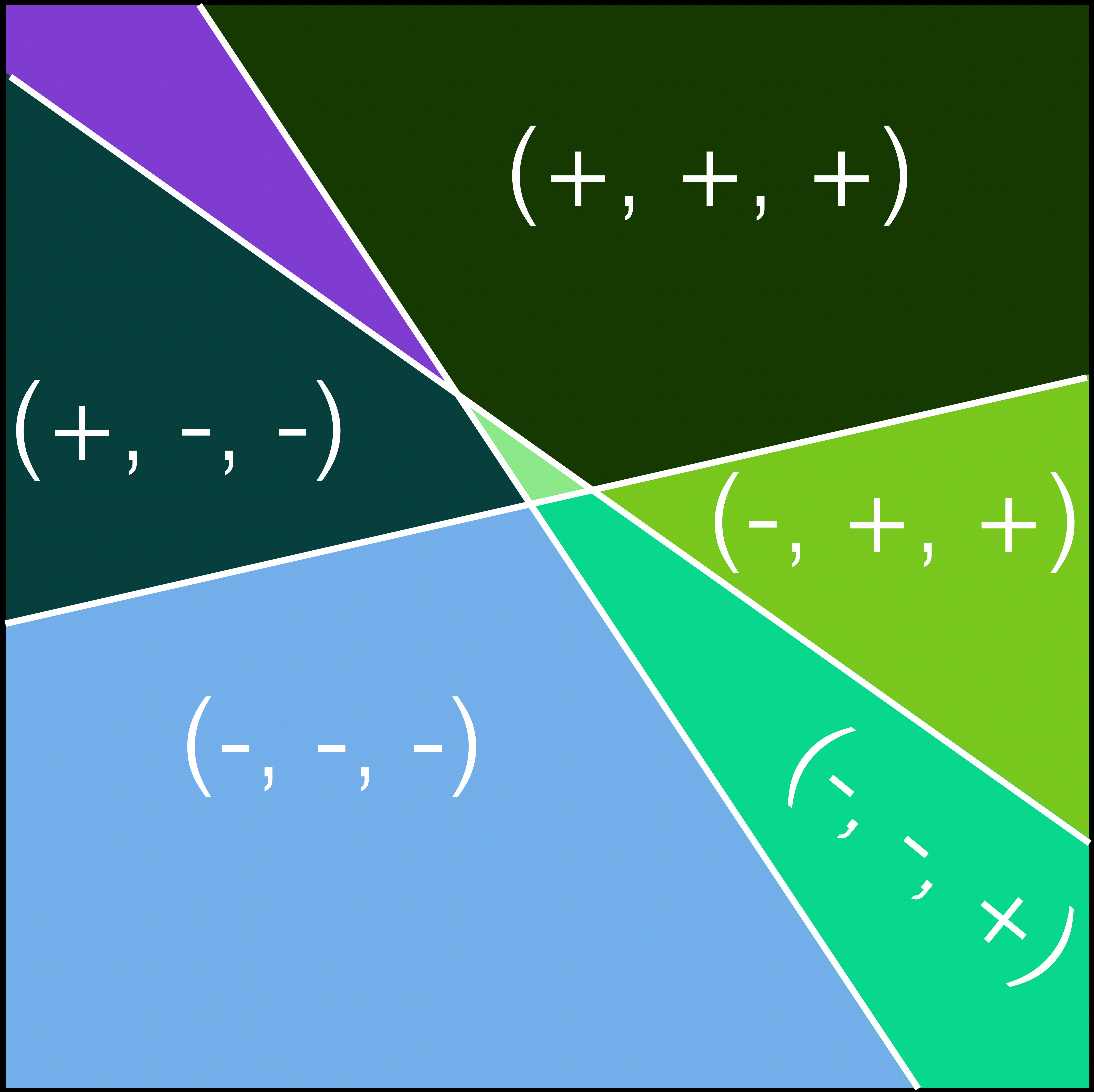


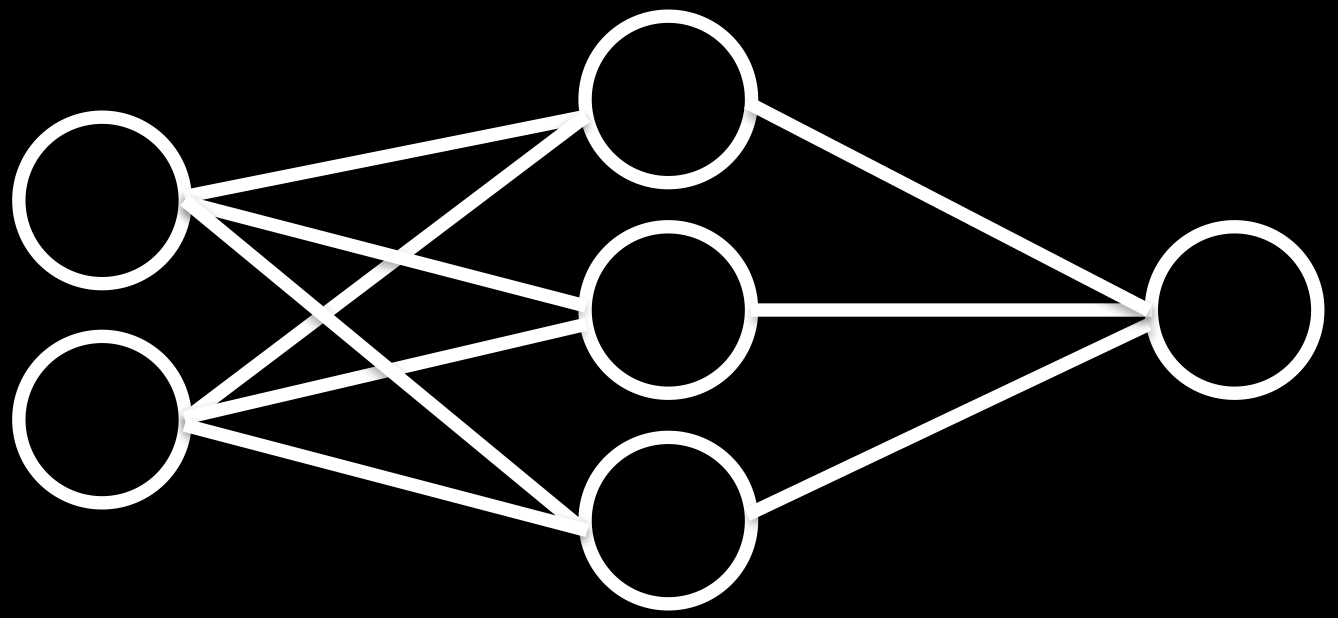


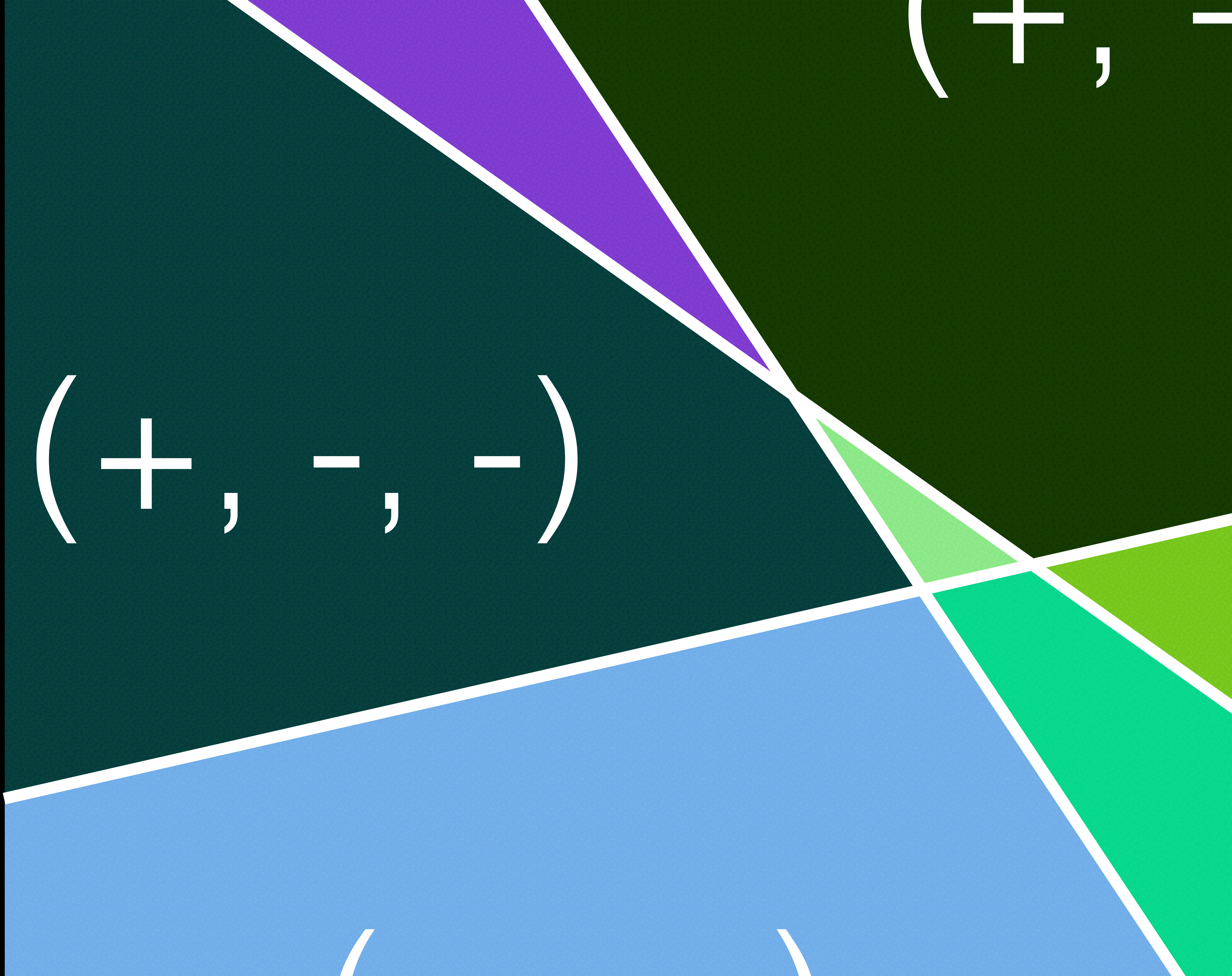


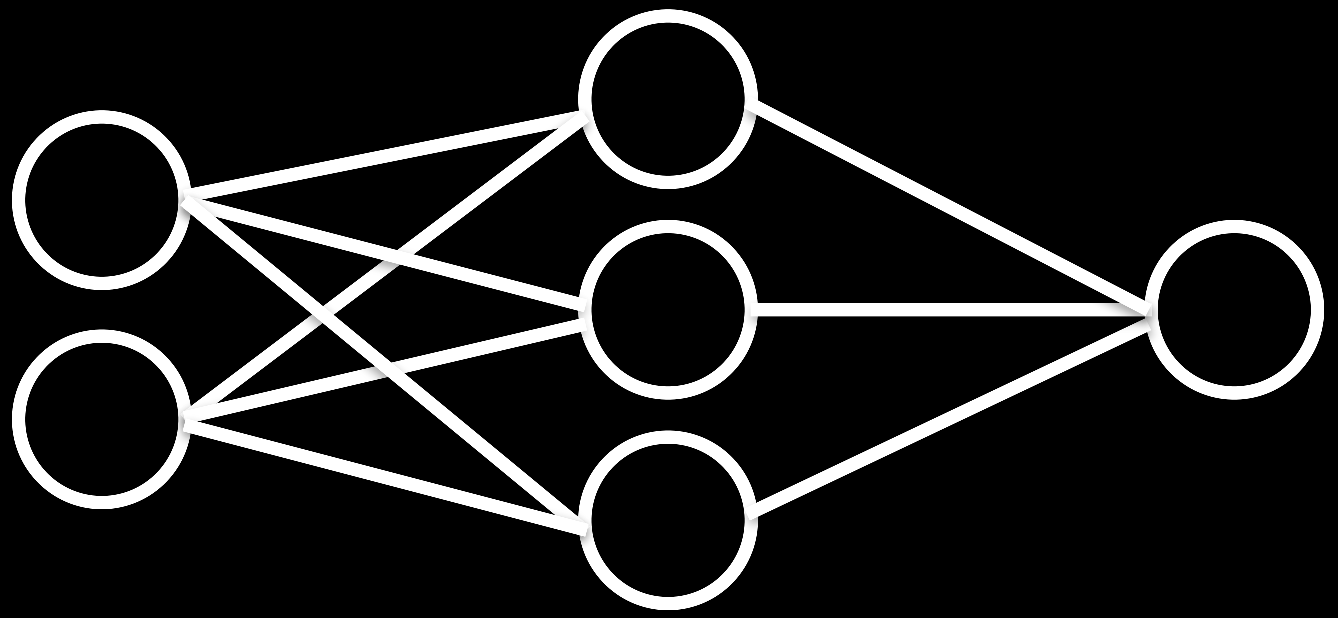


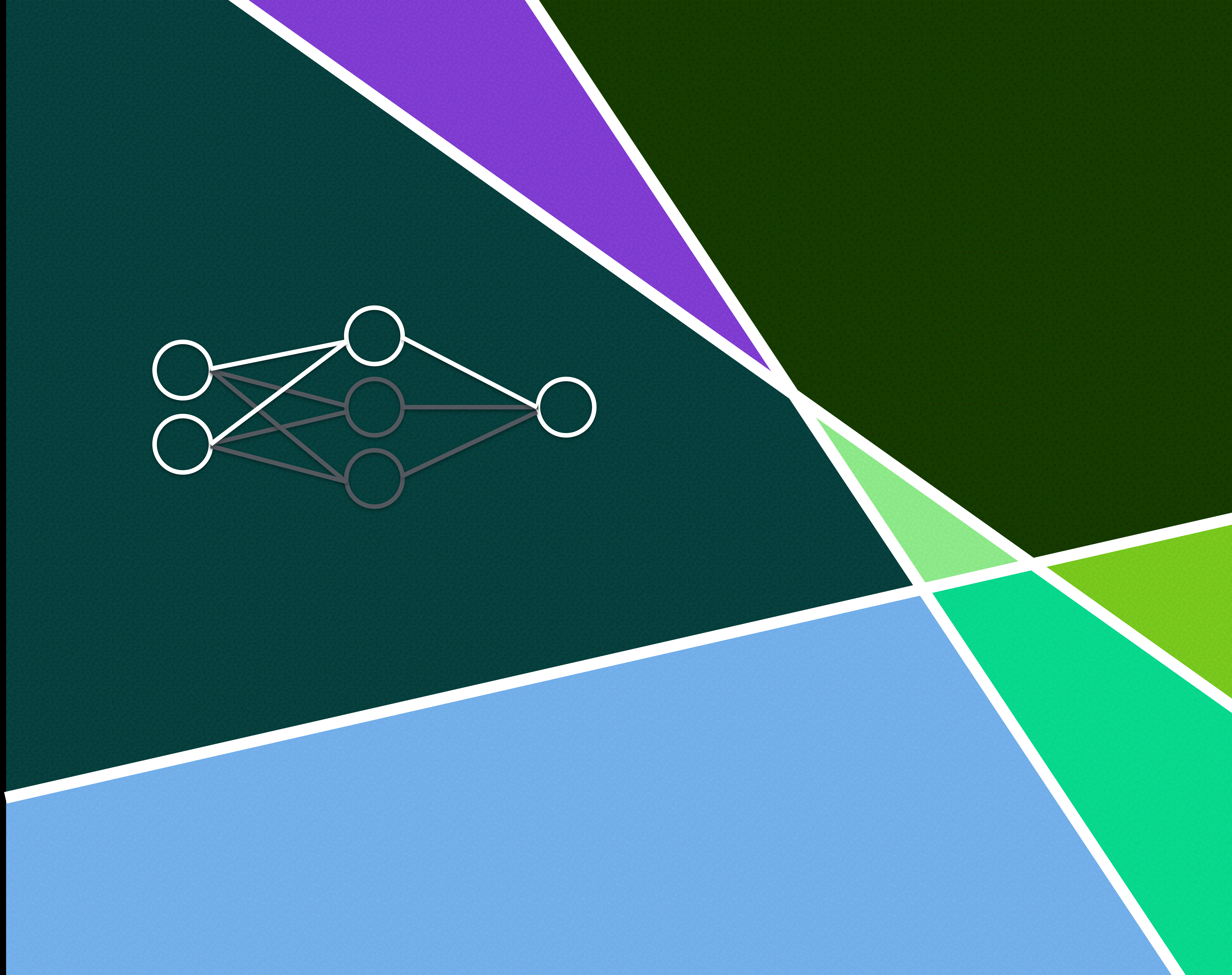


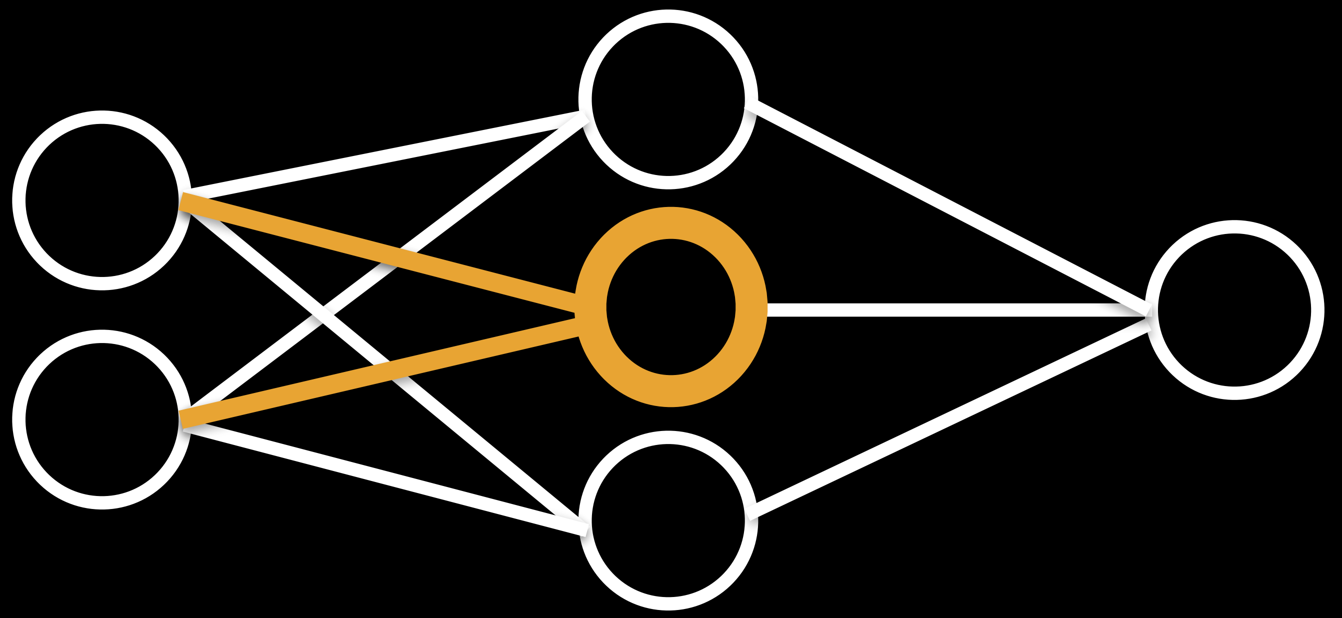


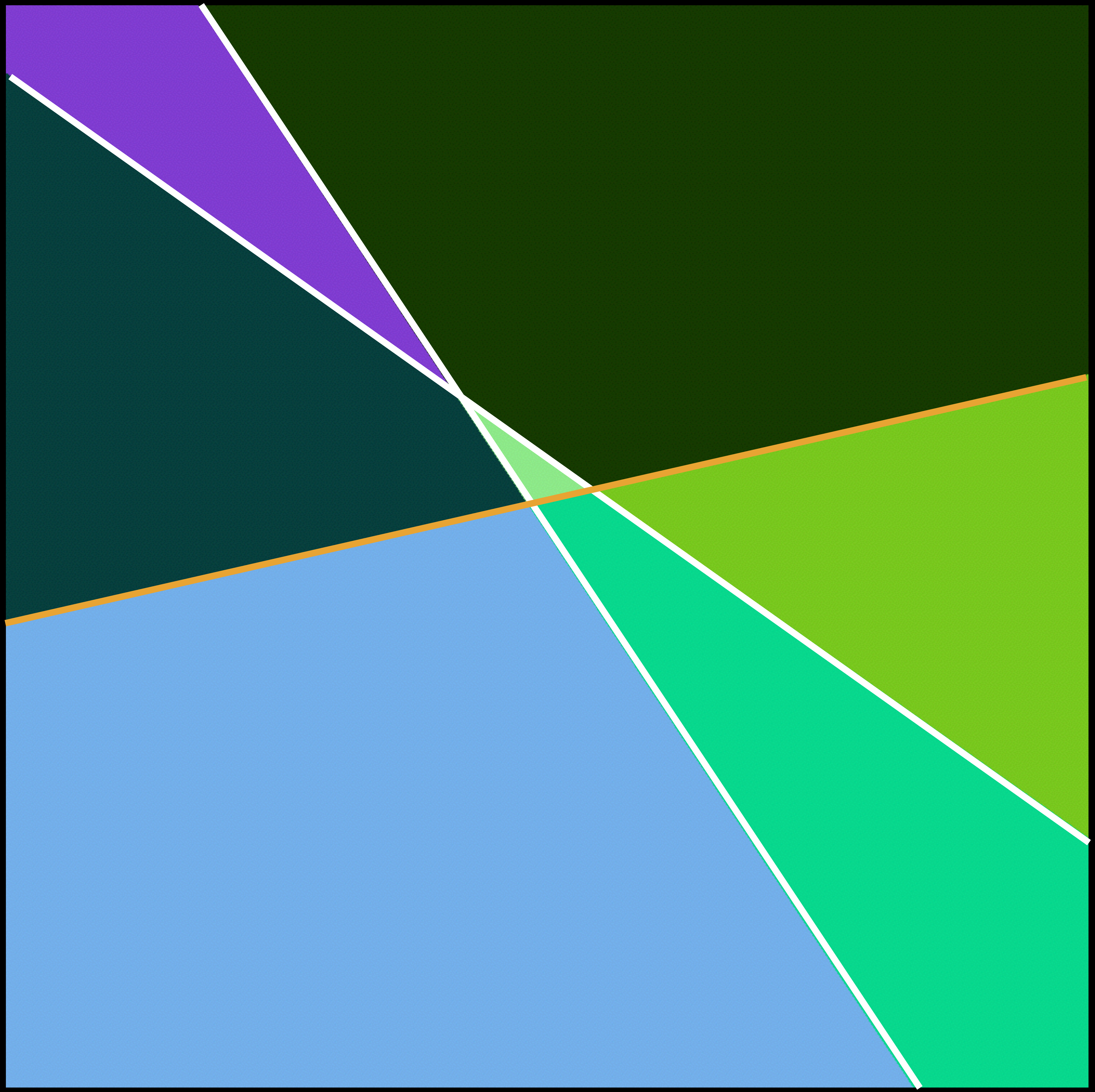


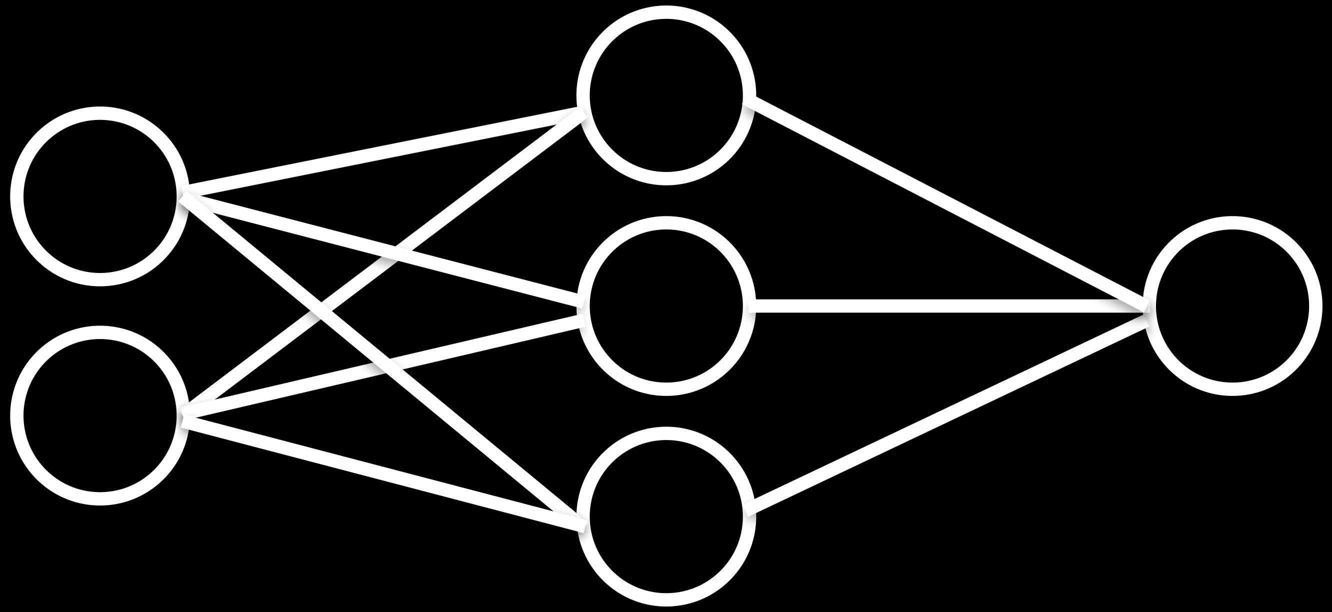


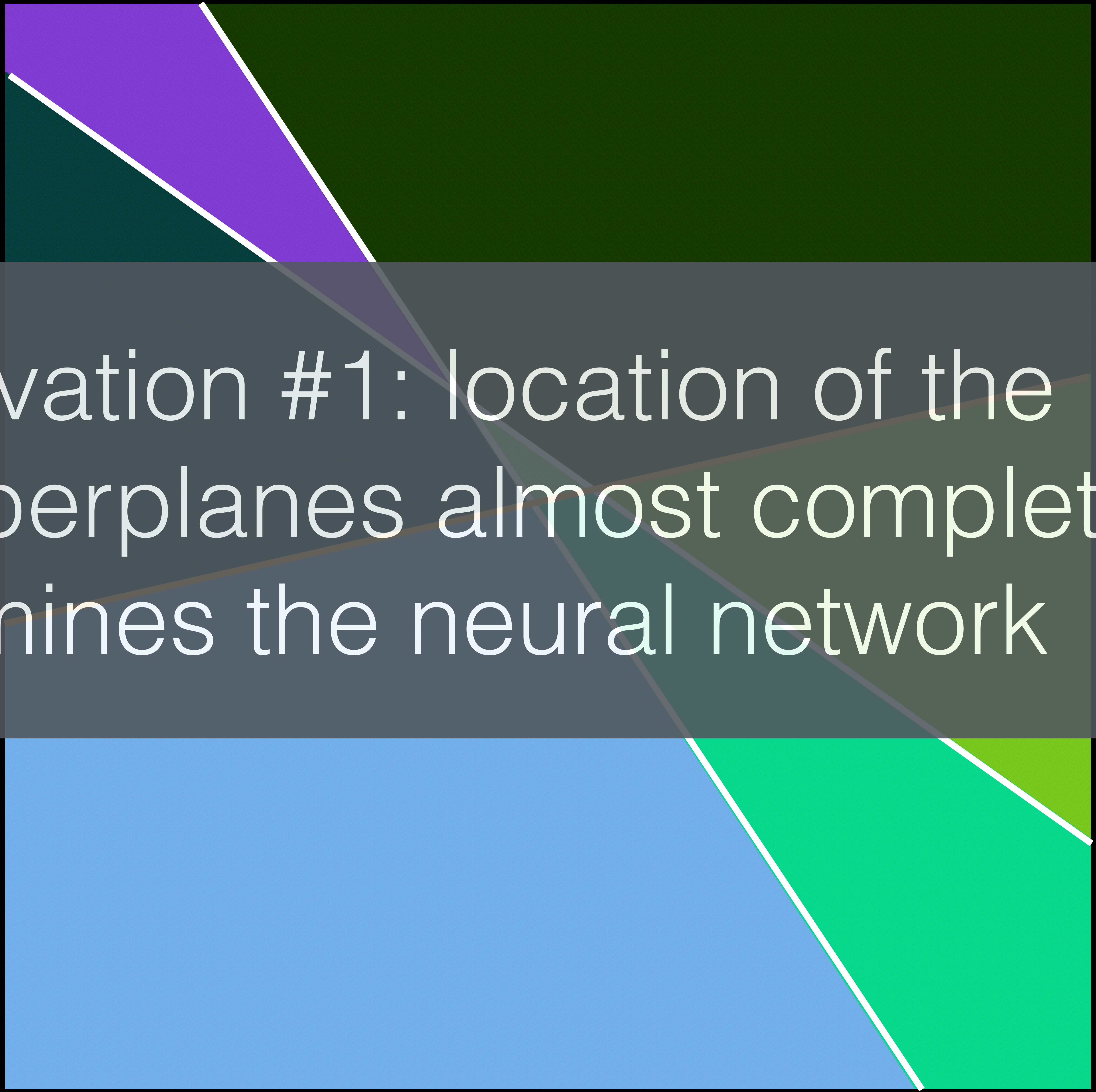




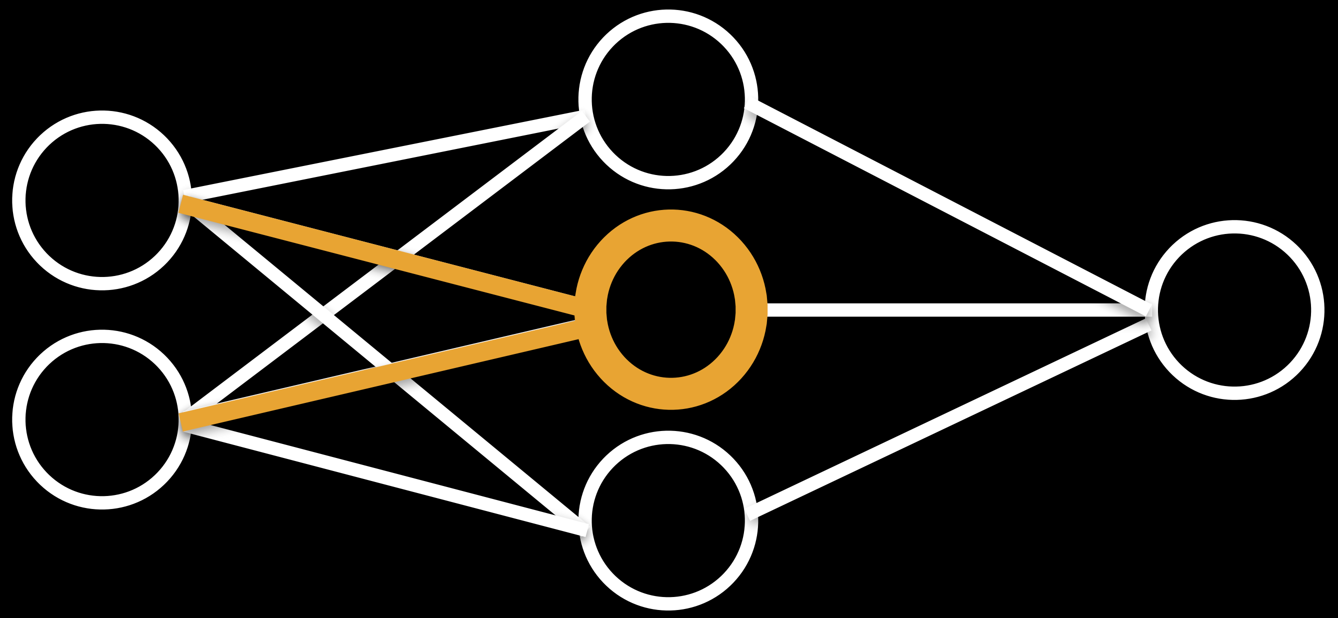


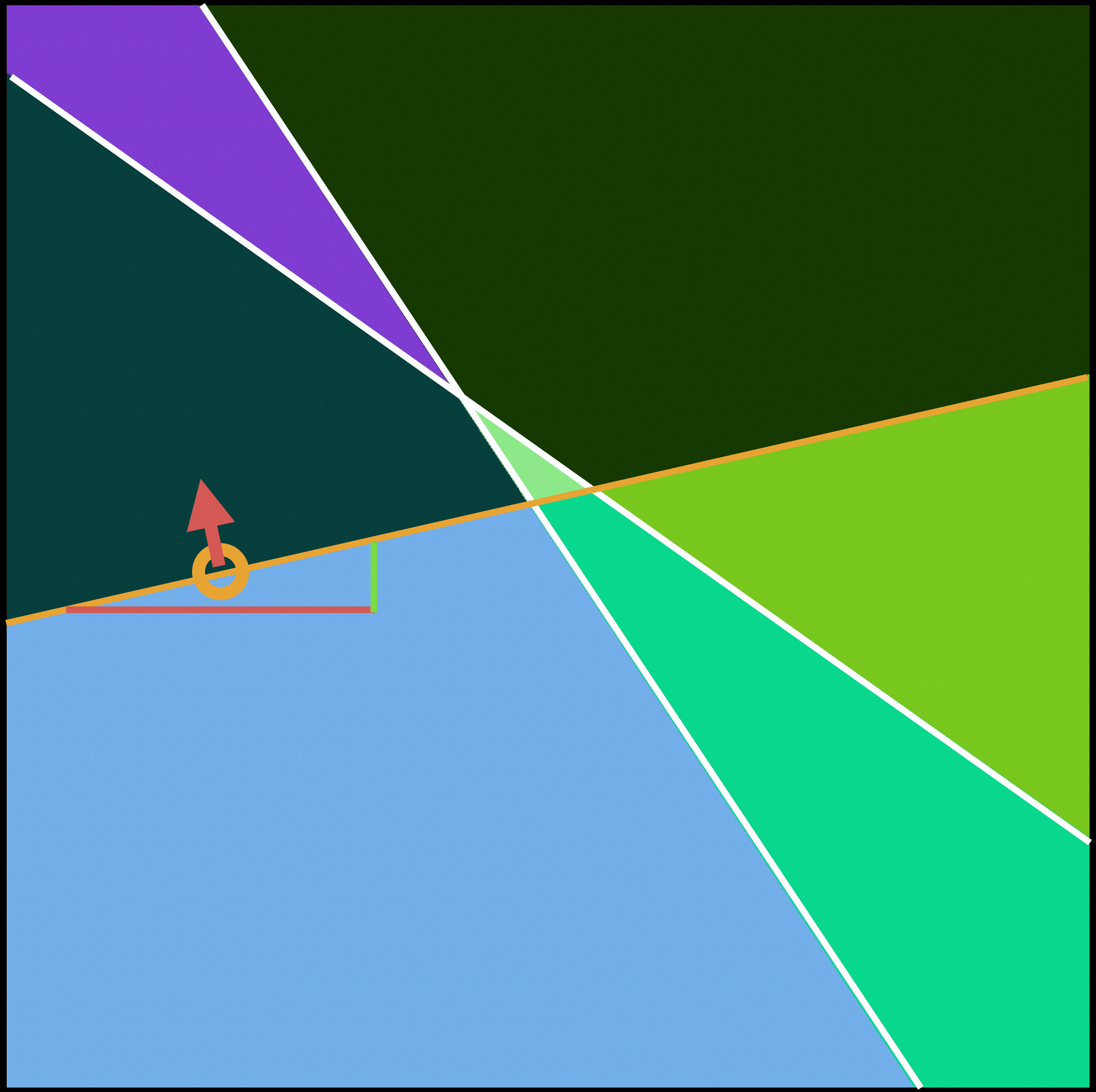




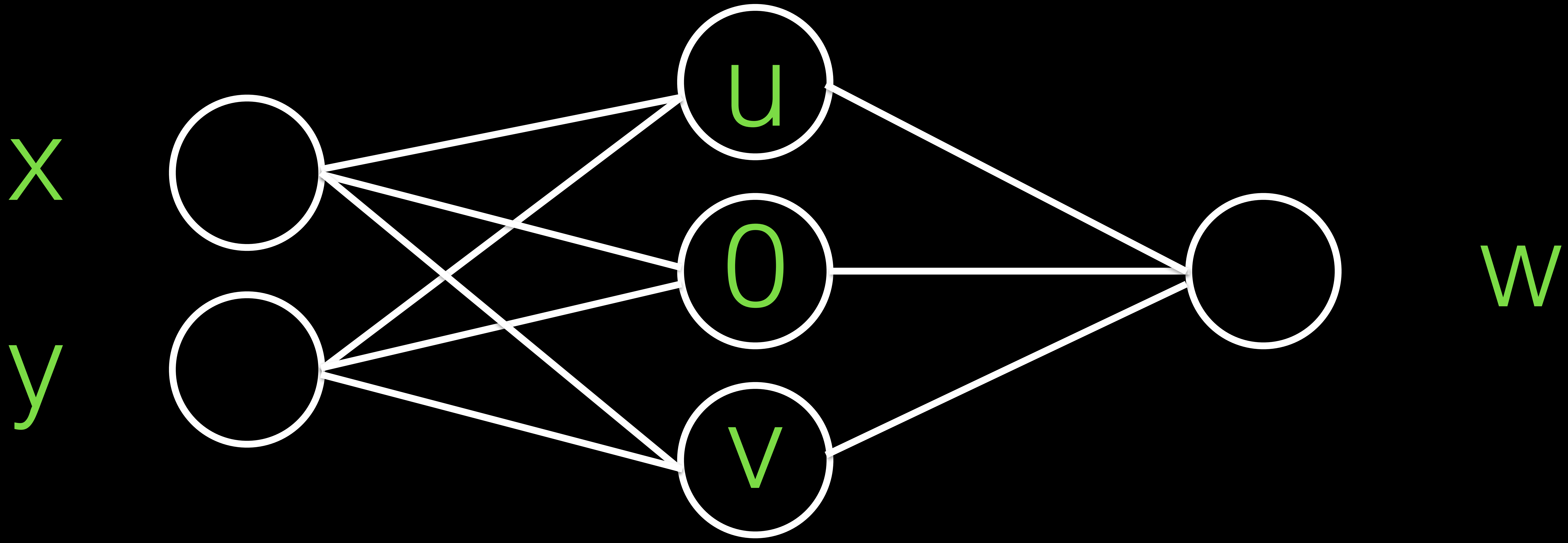


Observation #1: location of the critical hyperplanes almost completely determines the neural network

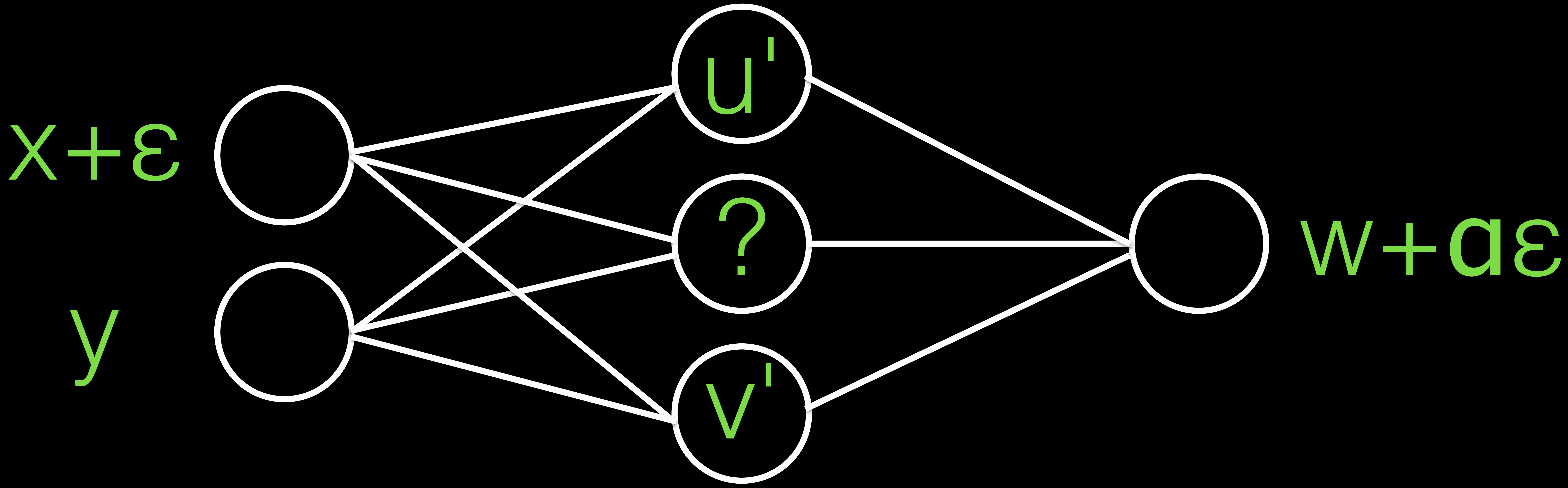




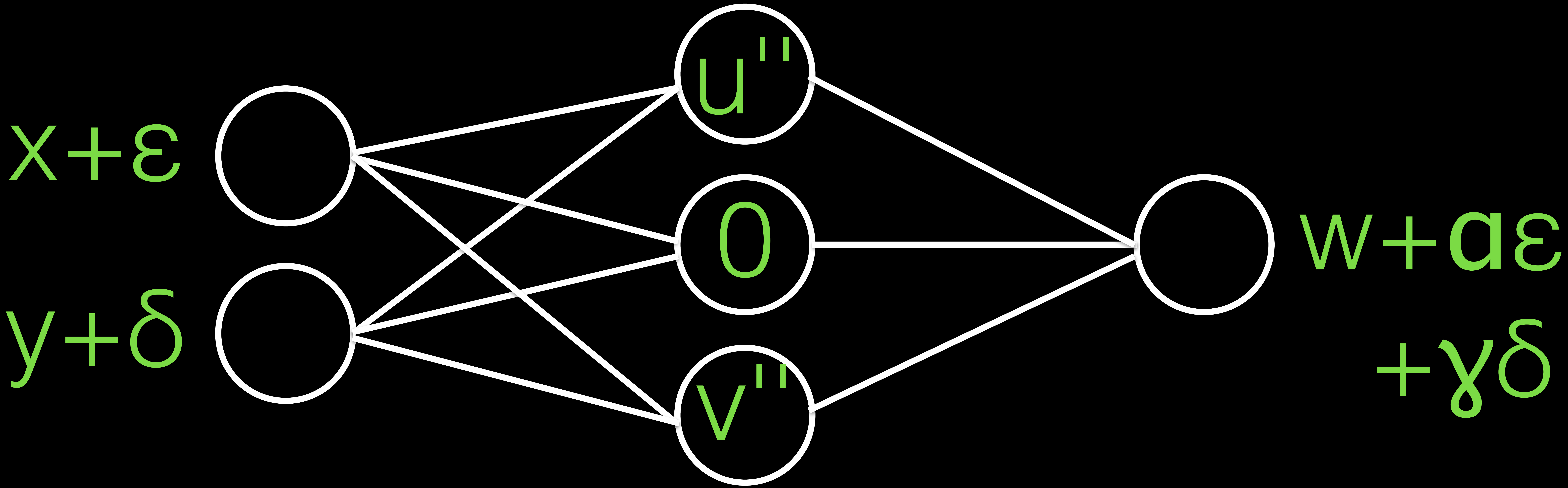
[MSDH19, JCB+20]



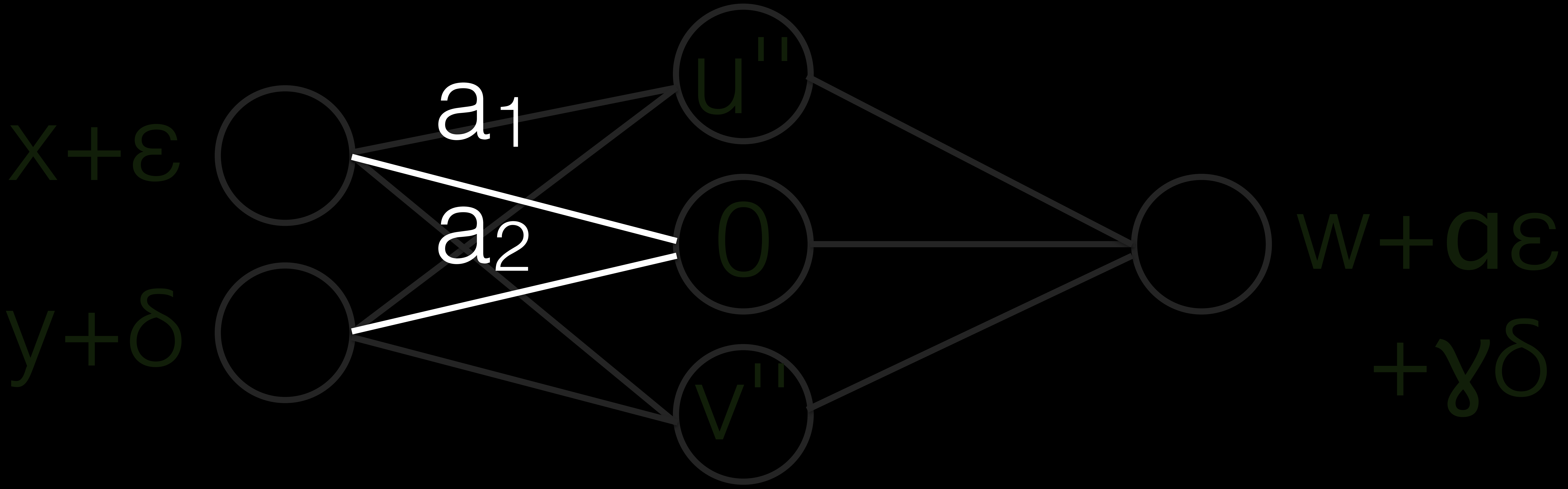
[MSDH19, JCB+20]



[MSDH19, JCB+20]

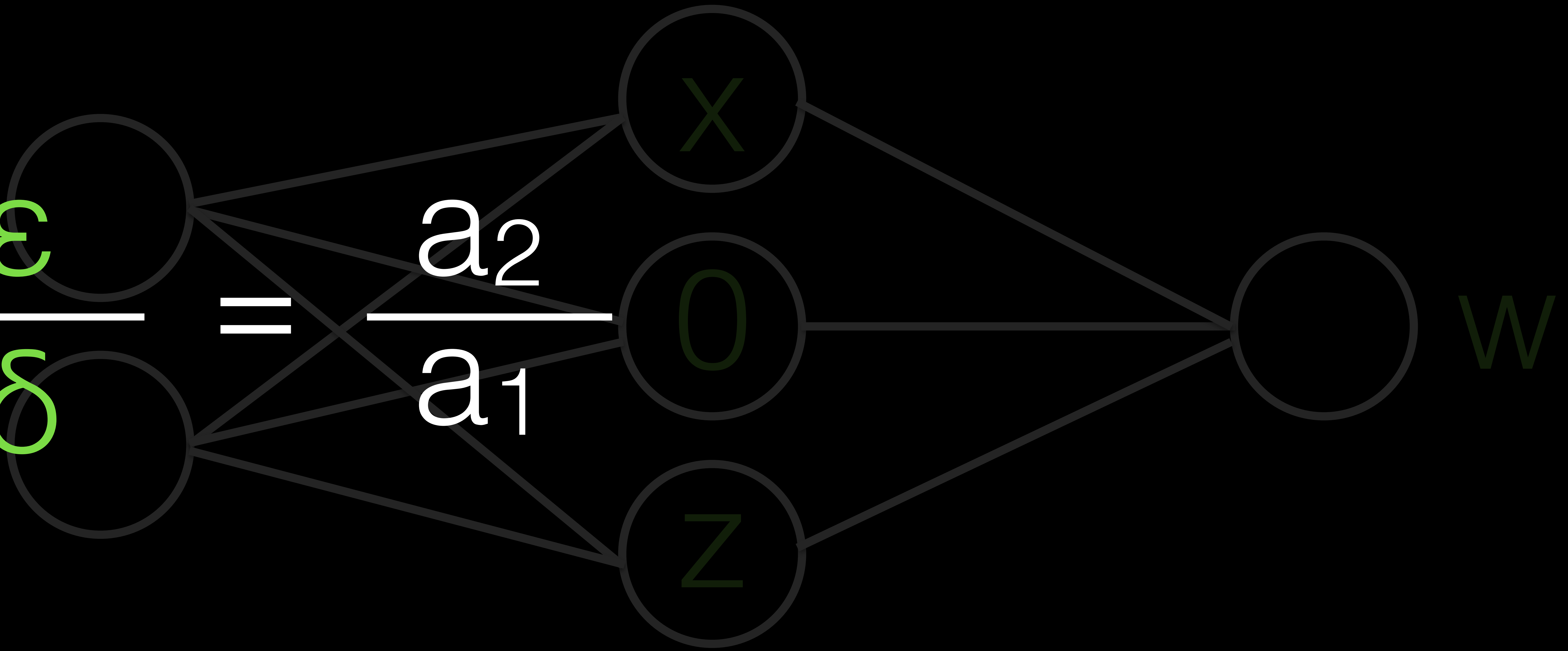


[MSDH19, JCB+20]

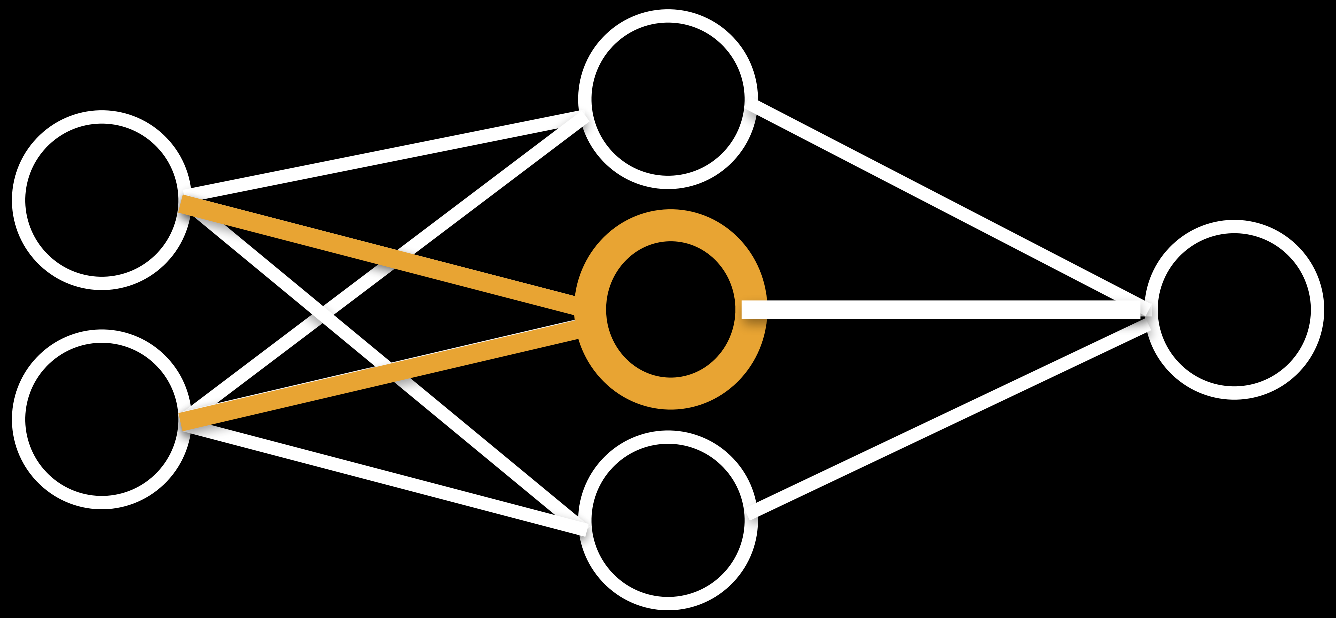


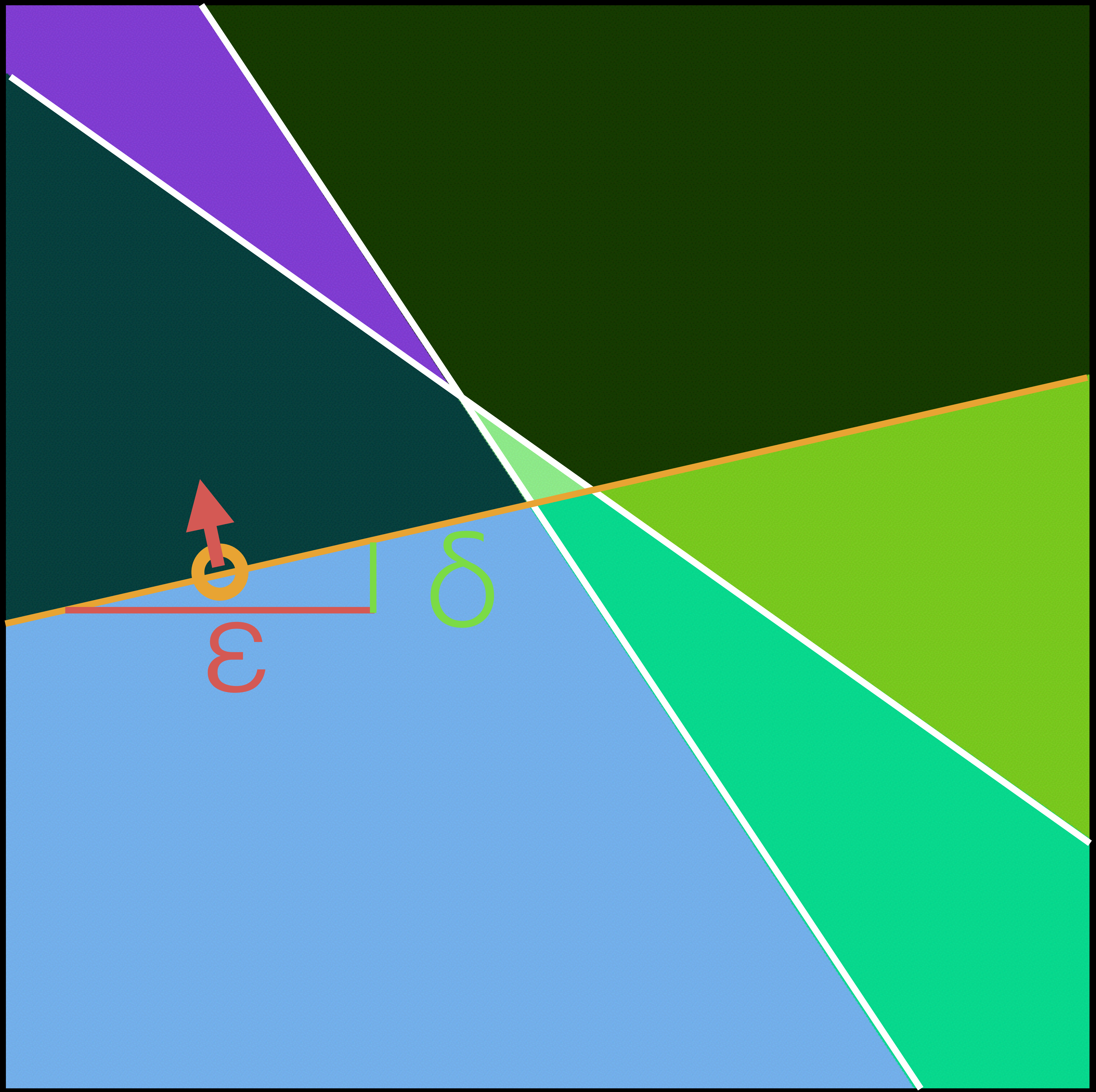
[MSDH19, JCB+20]

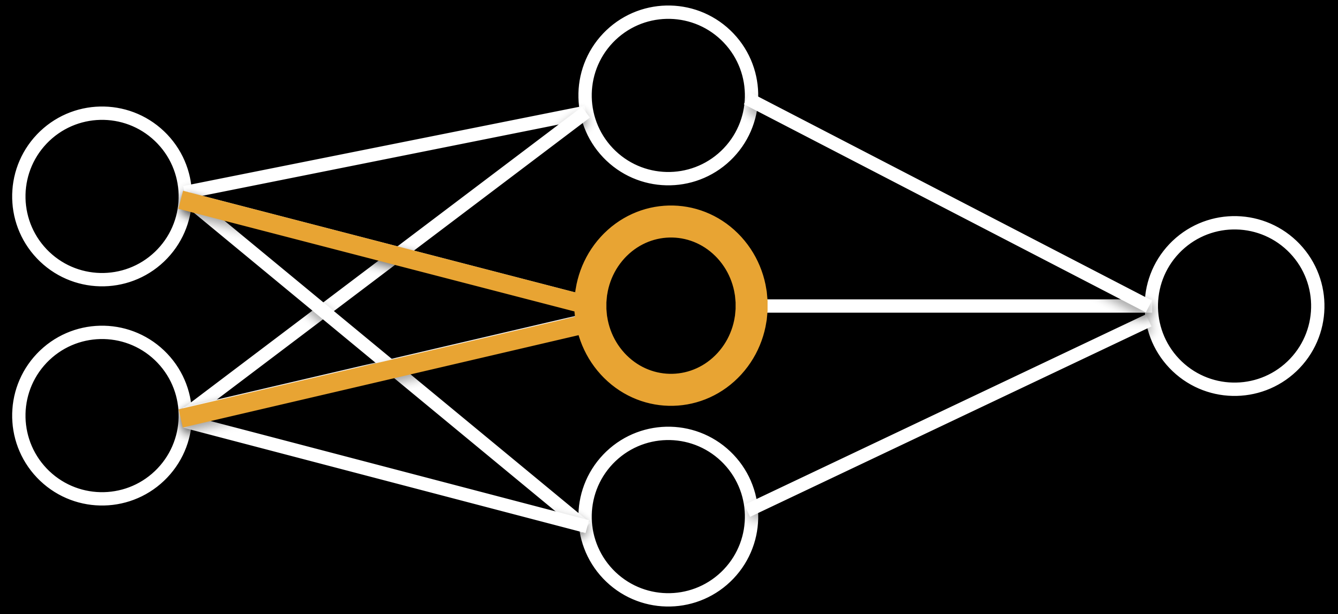
$$\frac{-\varepsilon}{\delta} = \frac{a_2}{a_1}$$

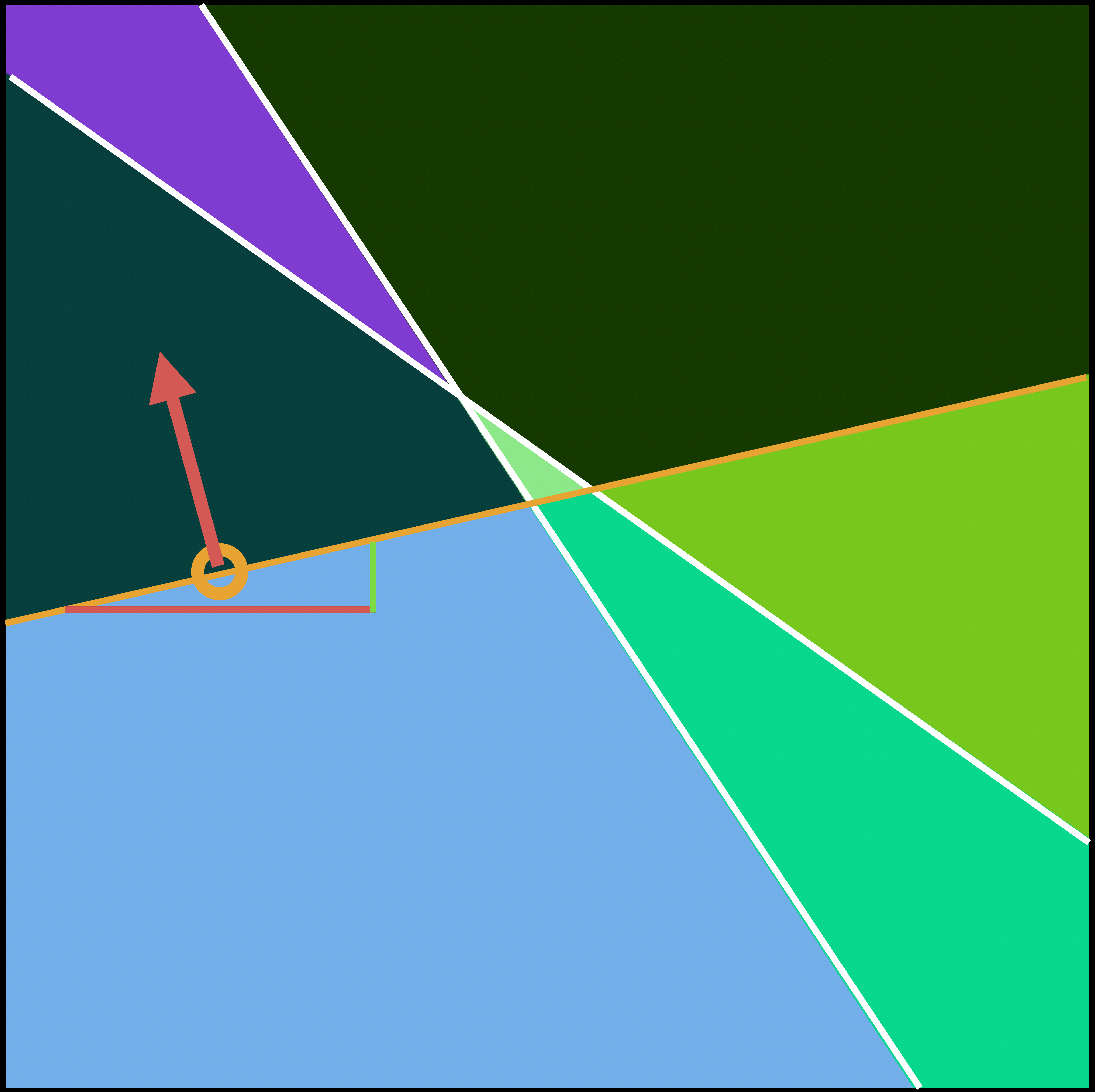


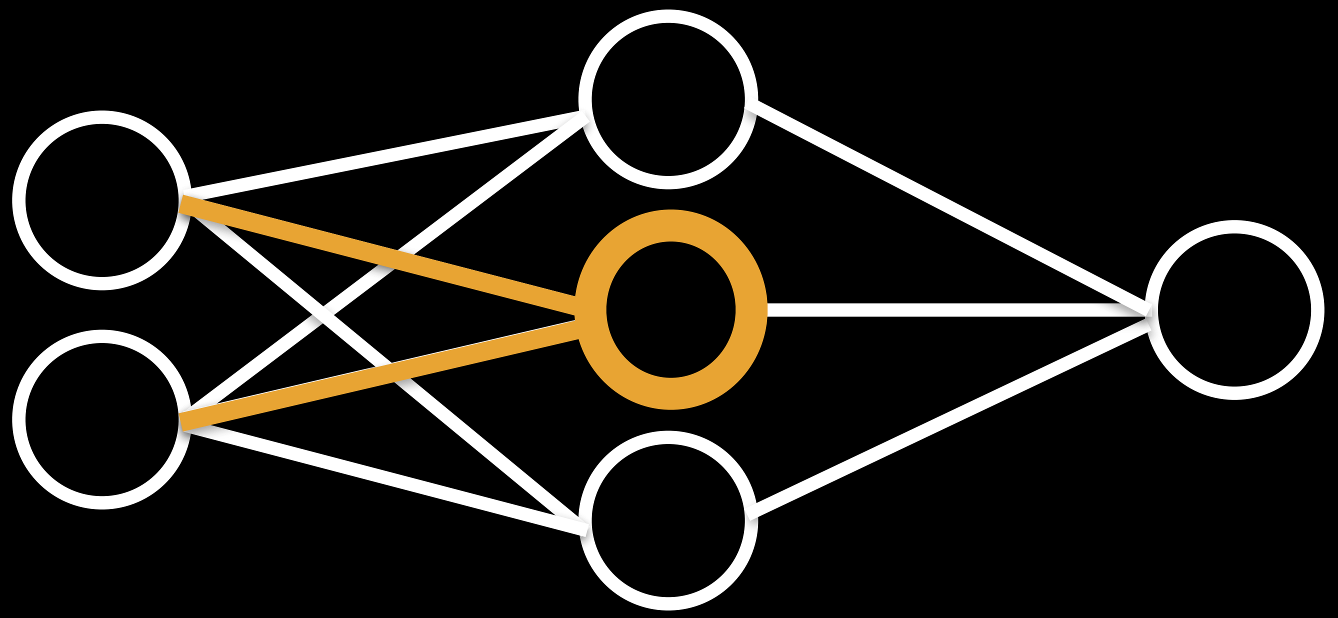
however….

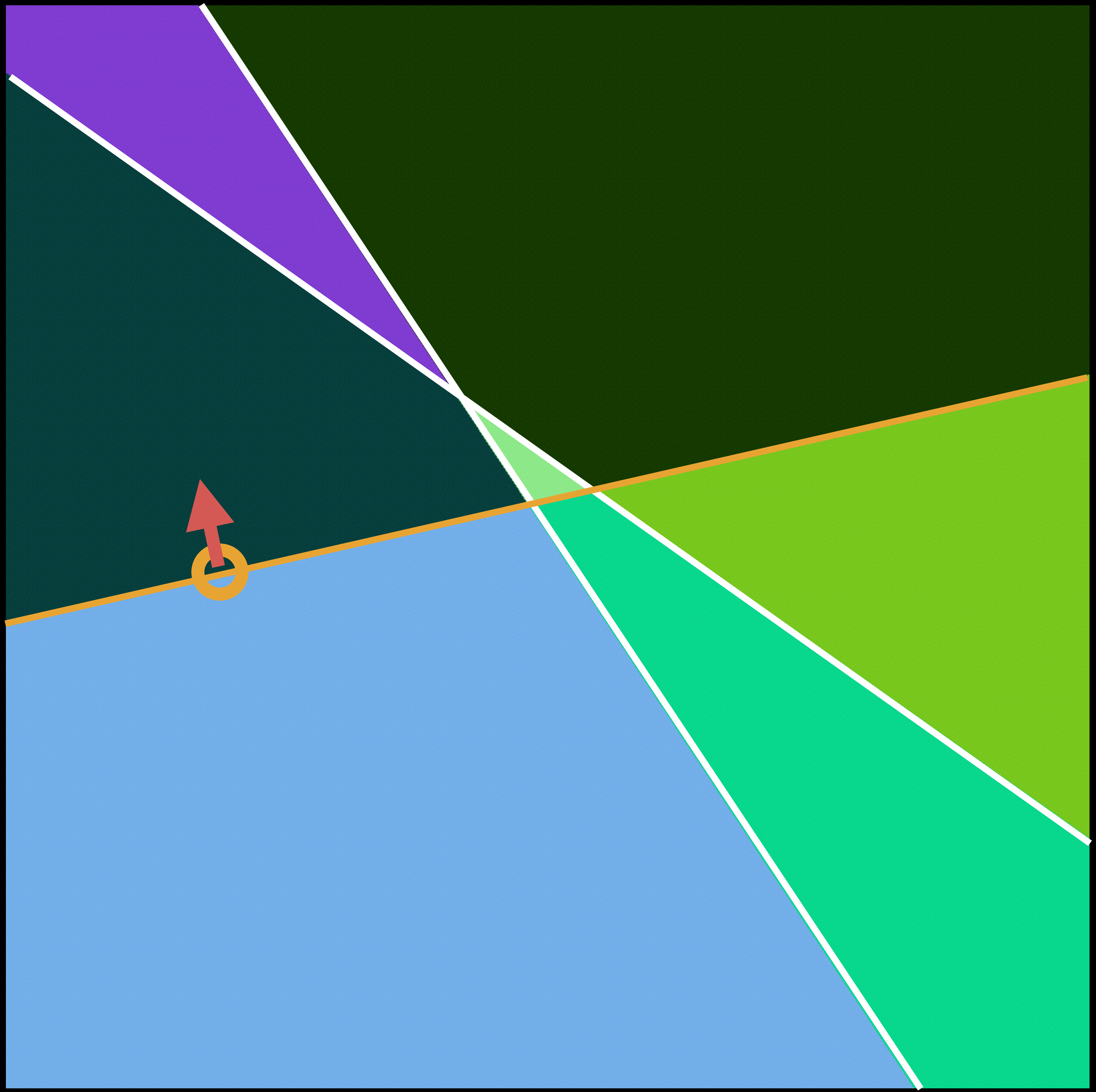


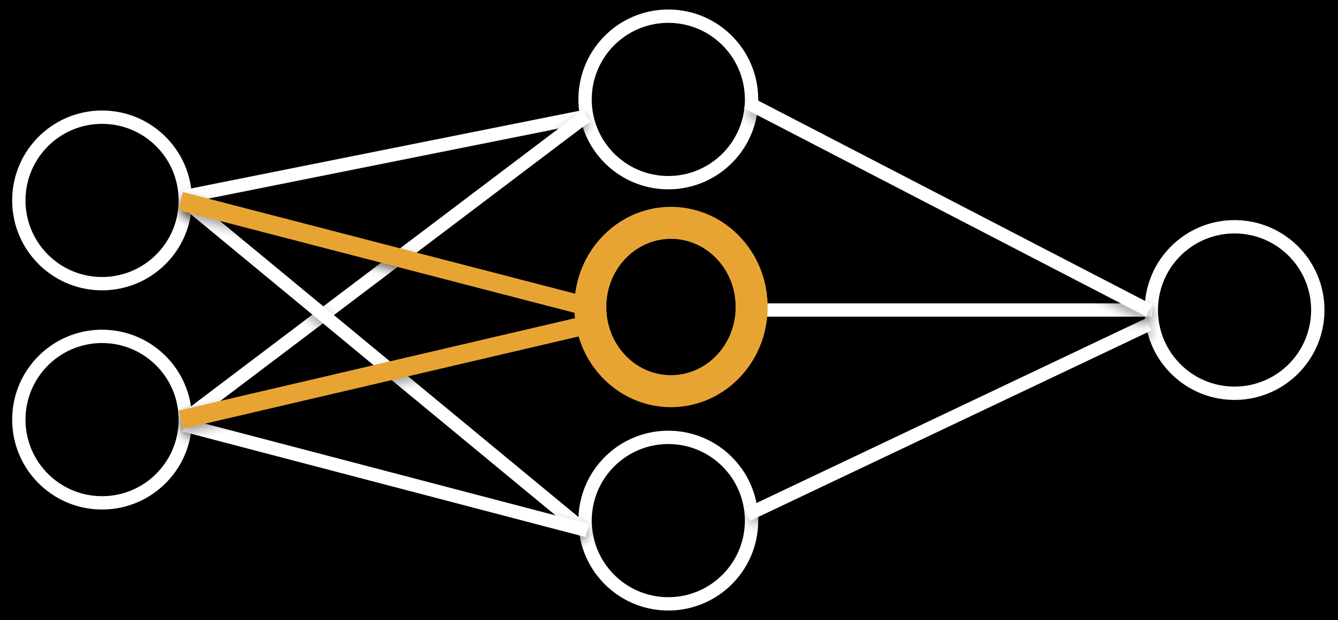

δ
ε

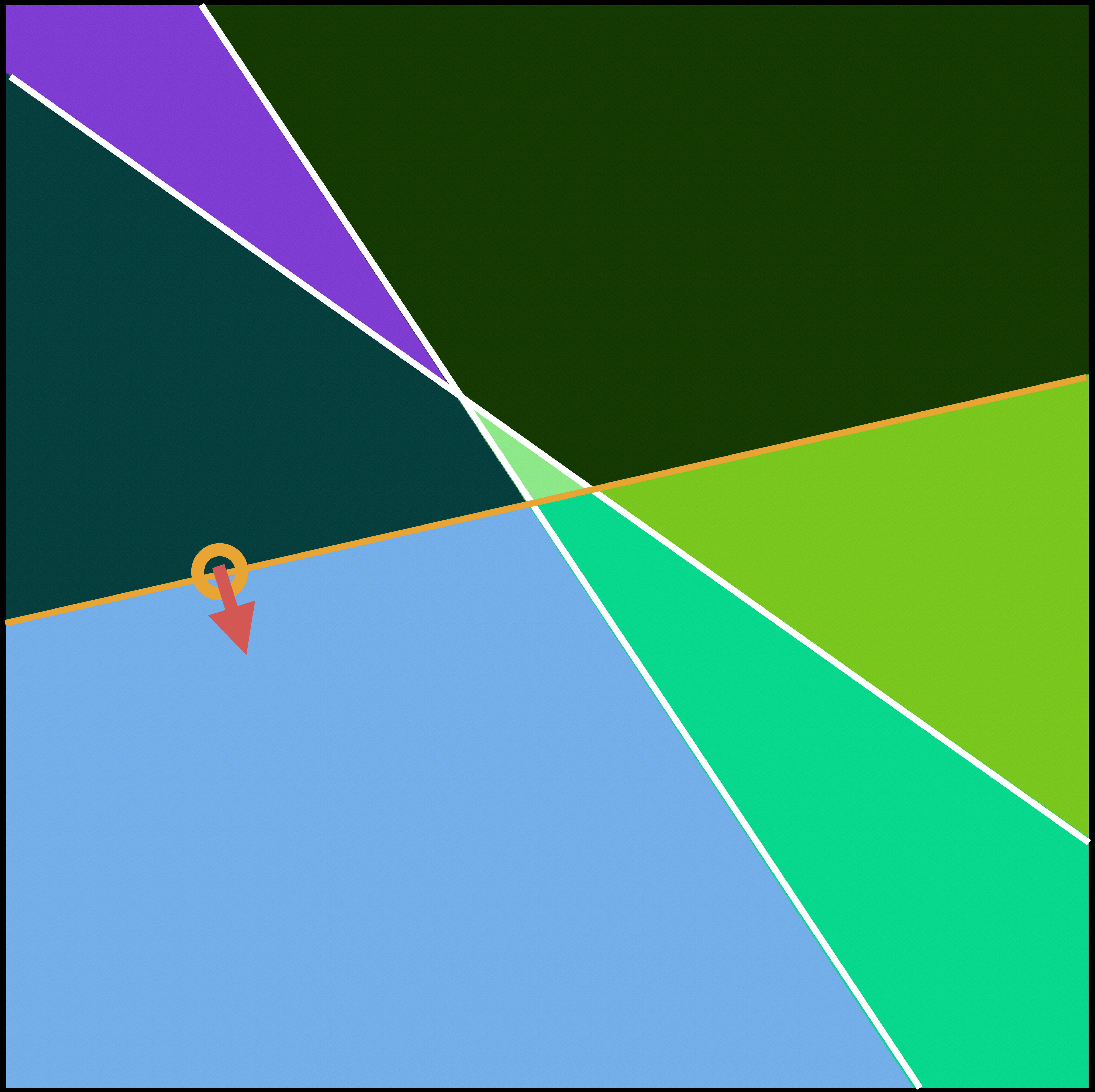


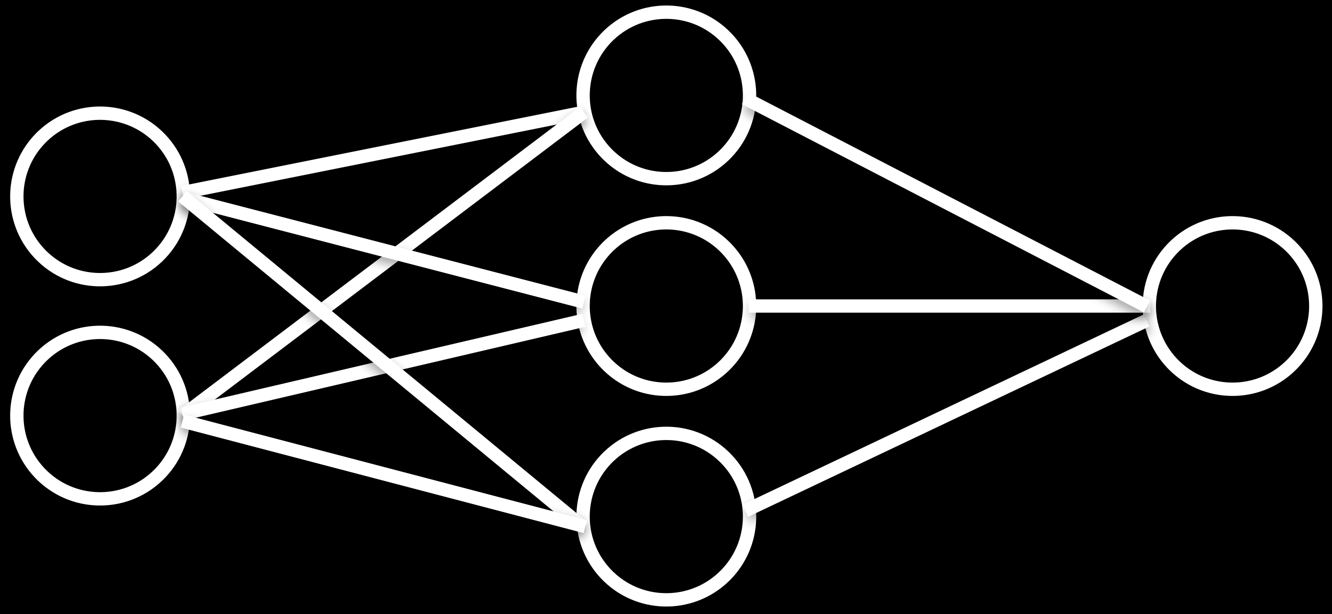


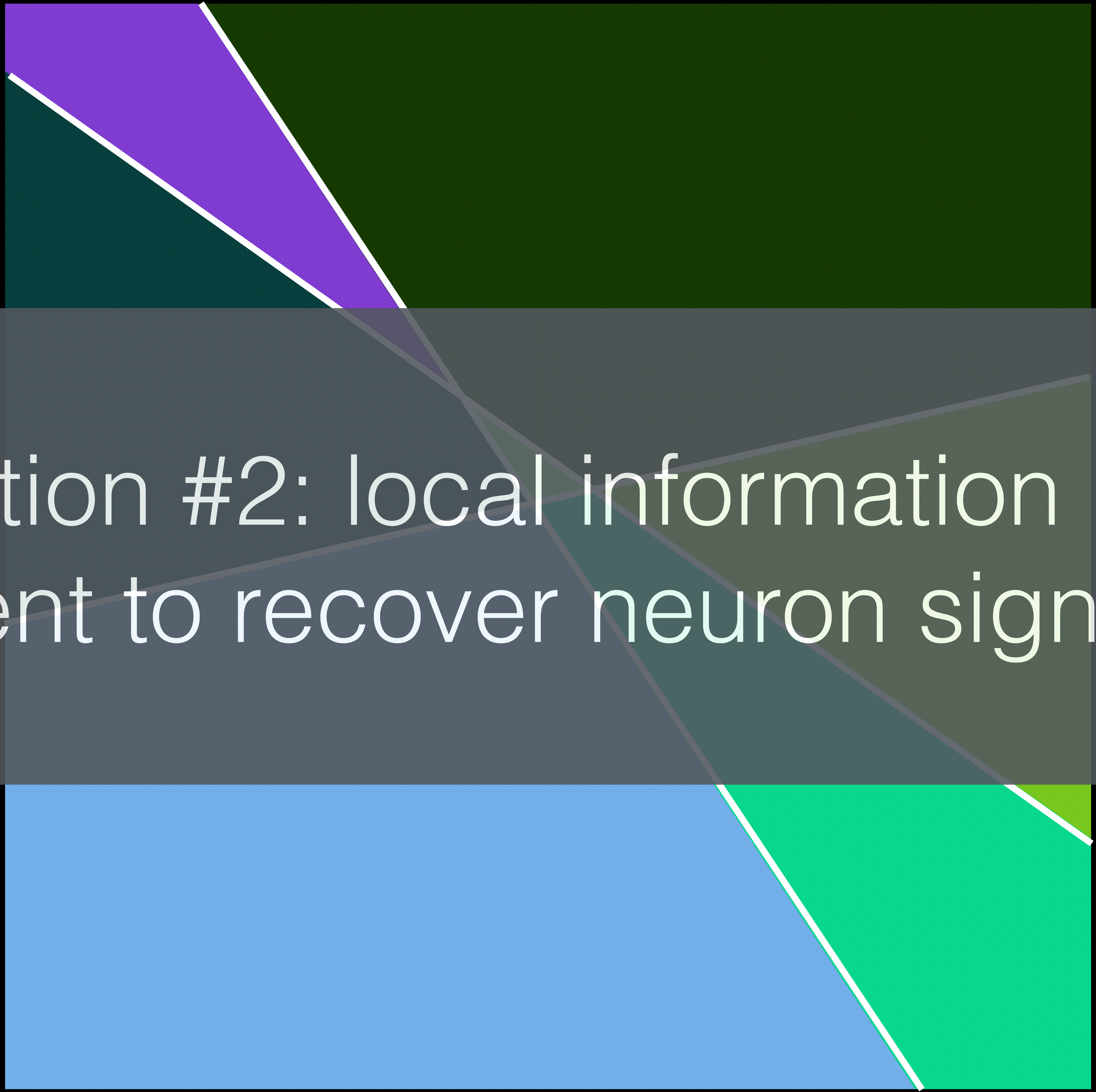




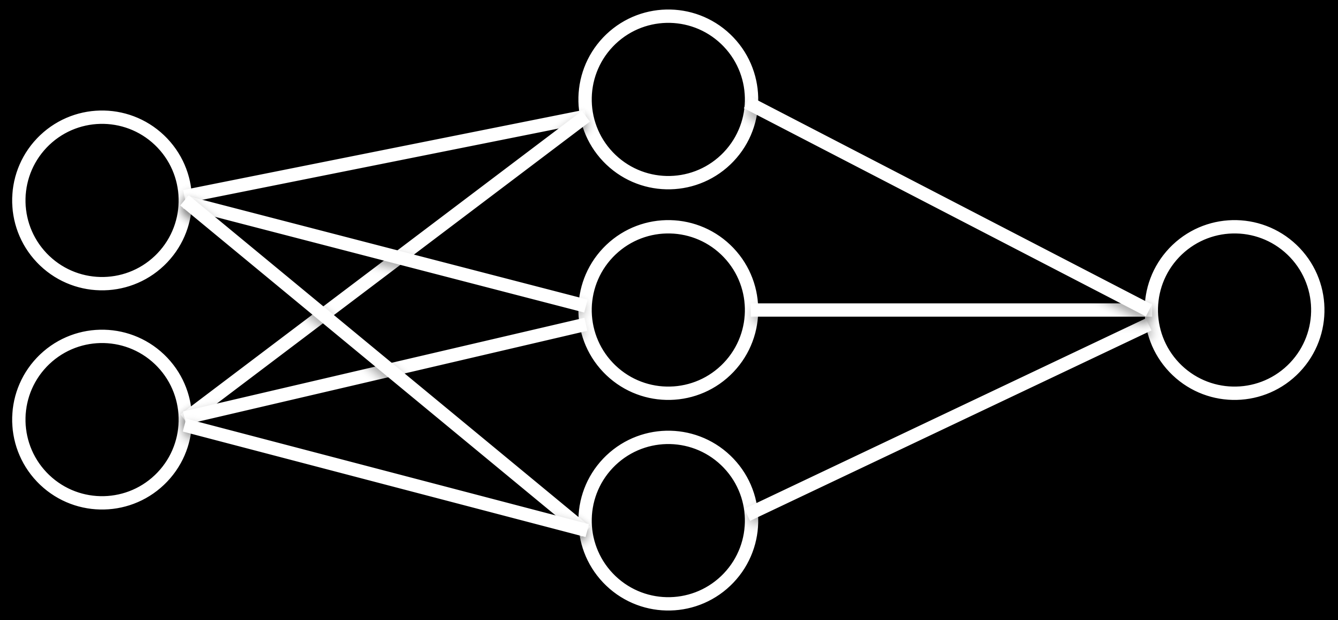


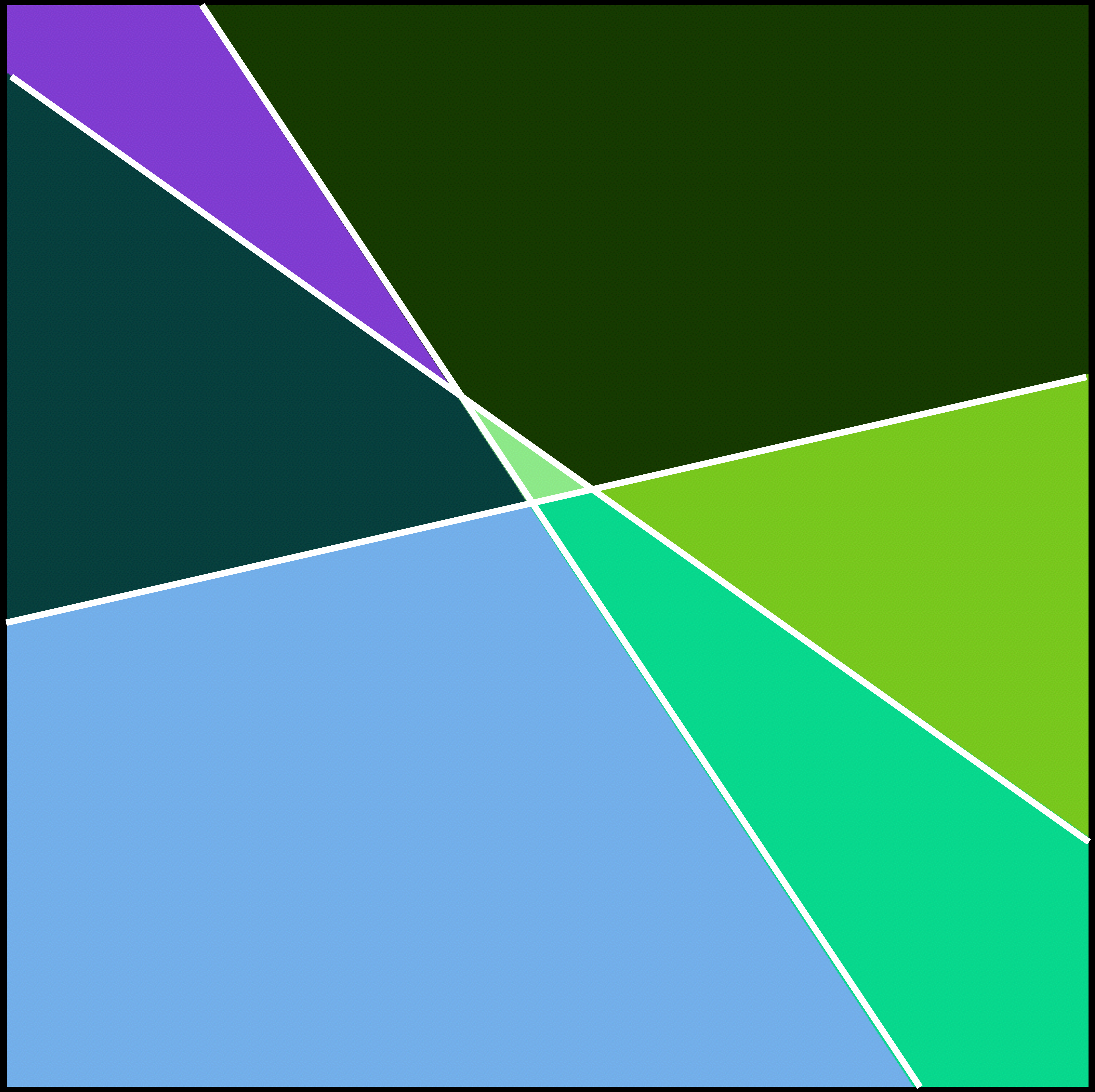


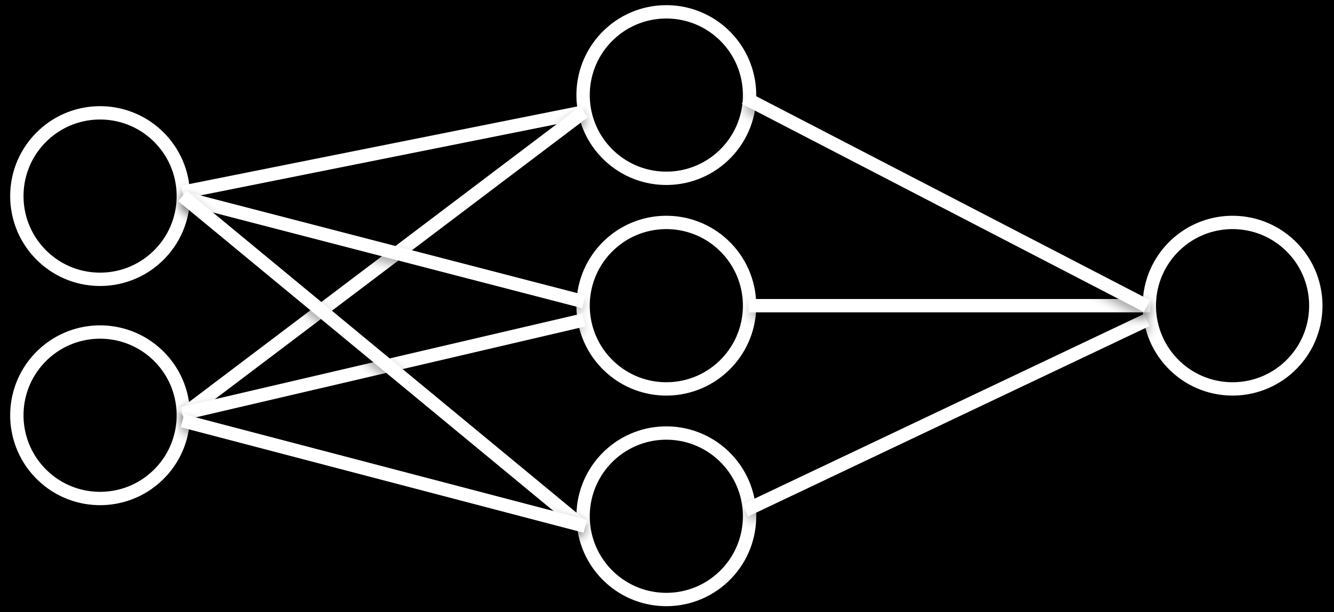


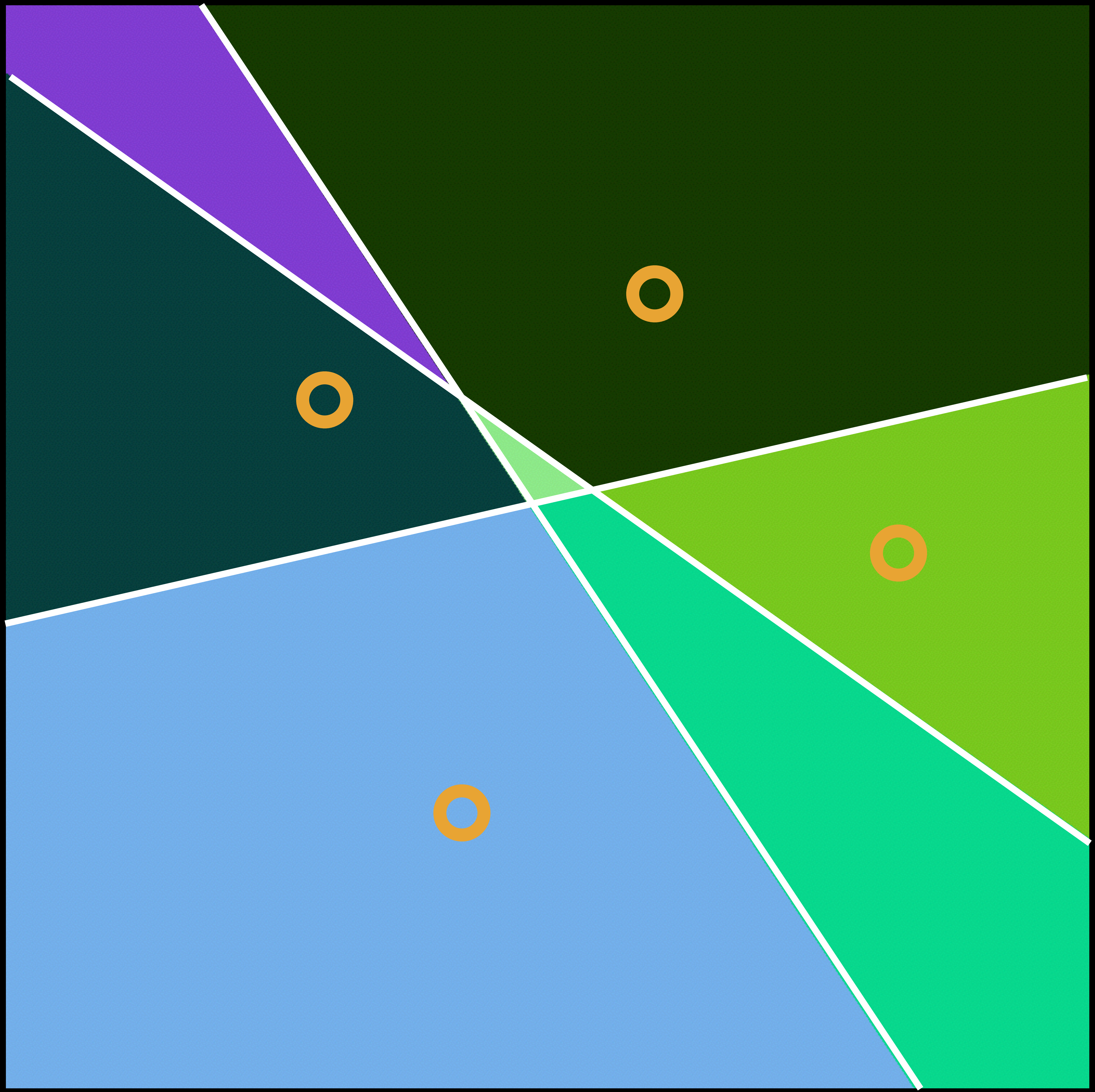


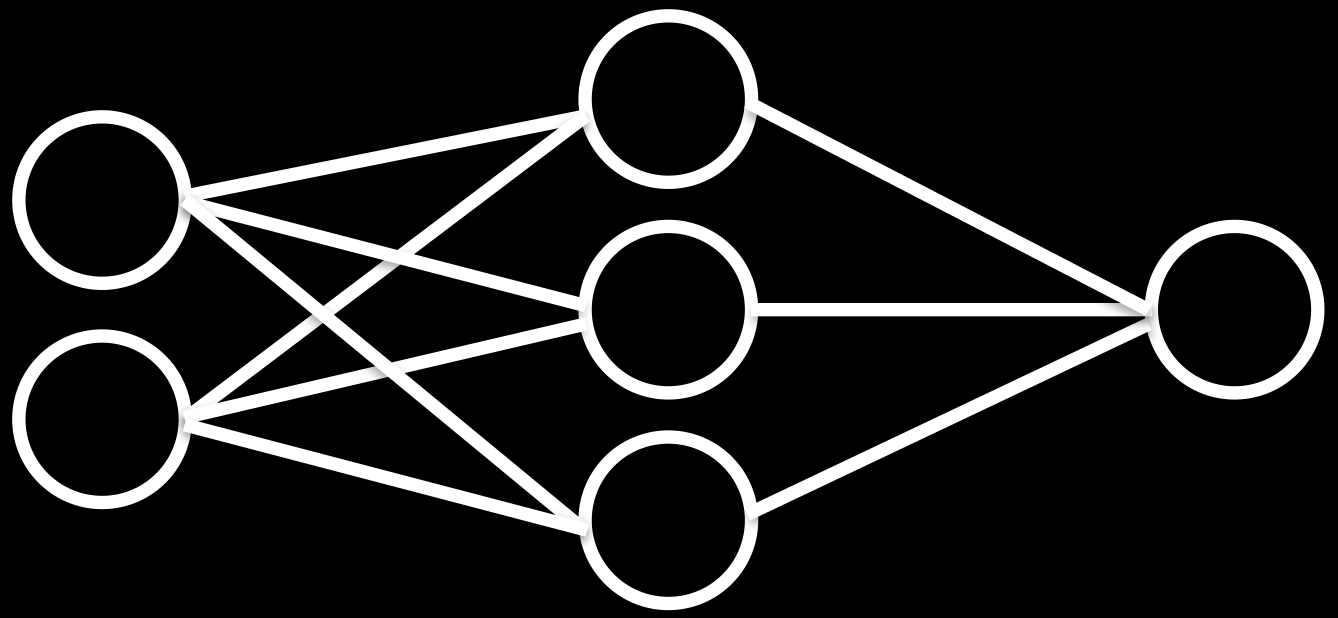
Observation #2: local information is insufficient to recover neuron signs

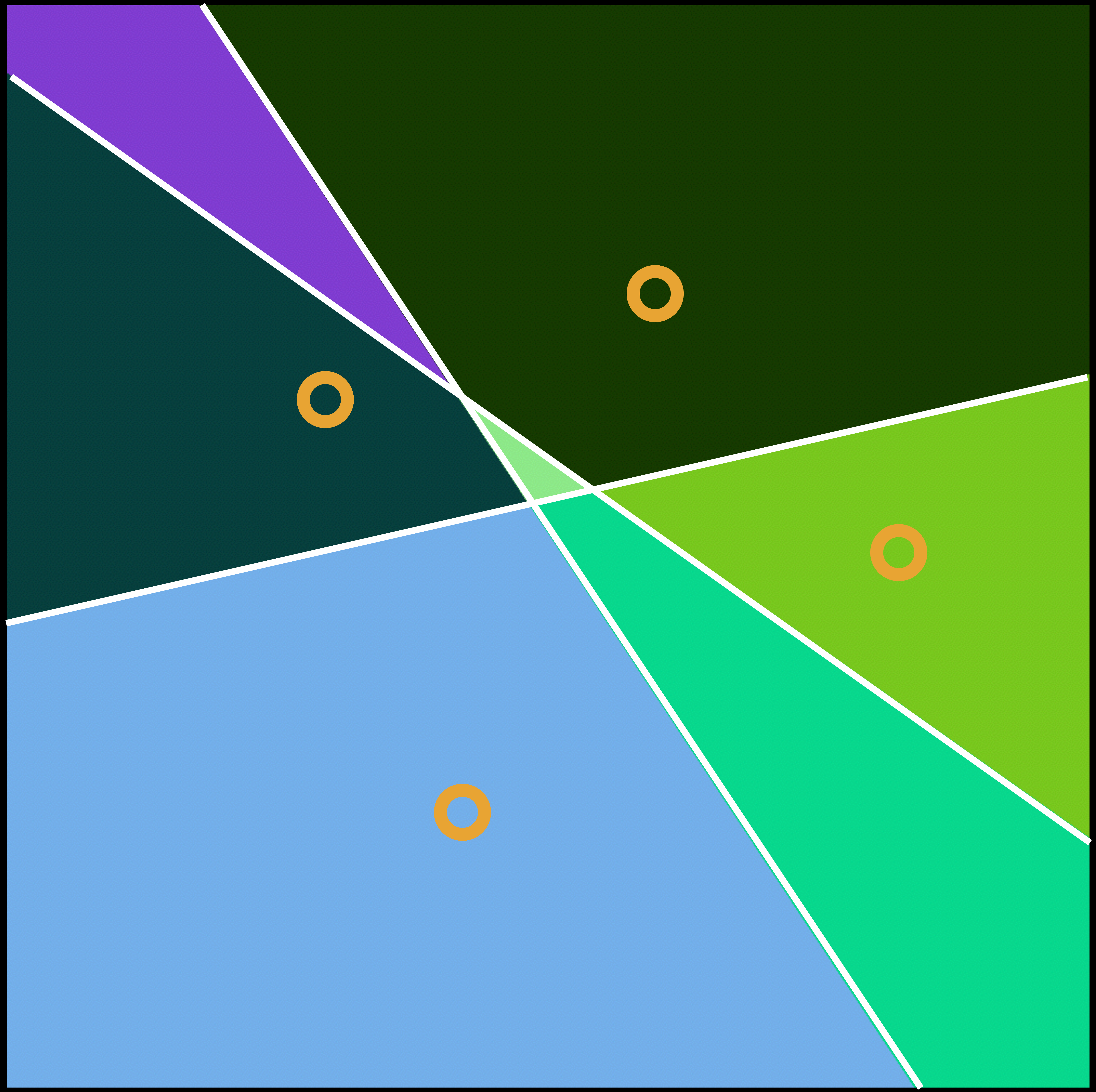




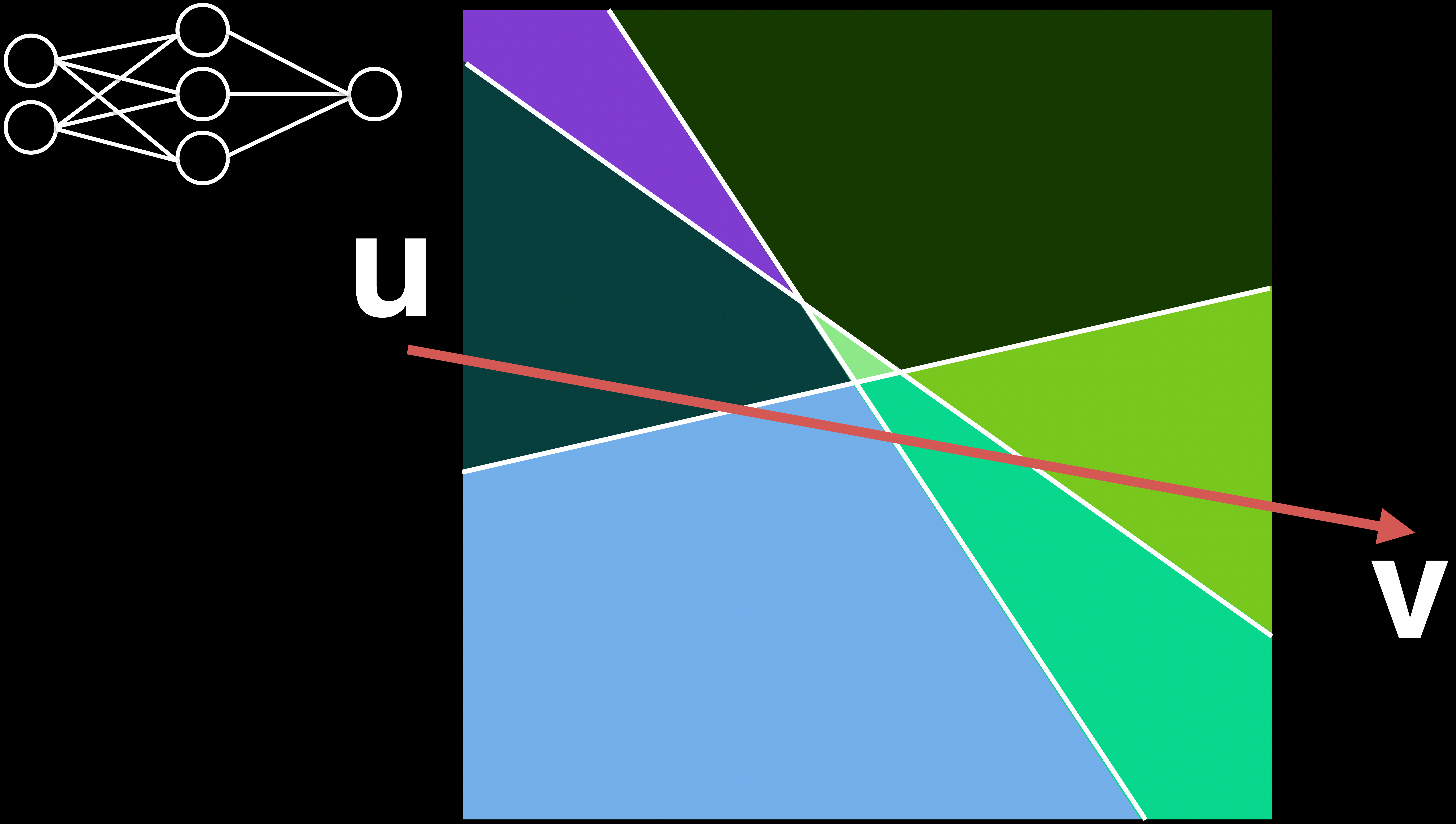


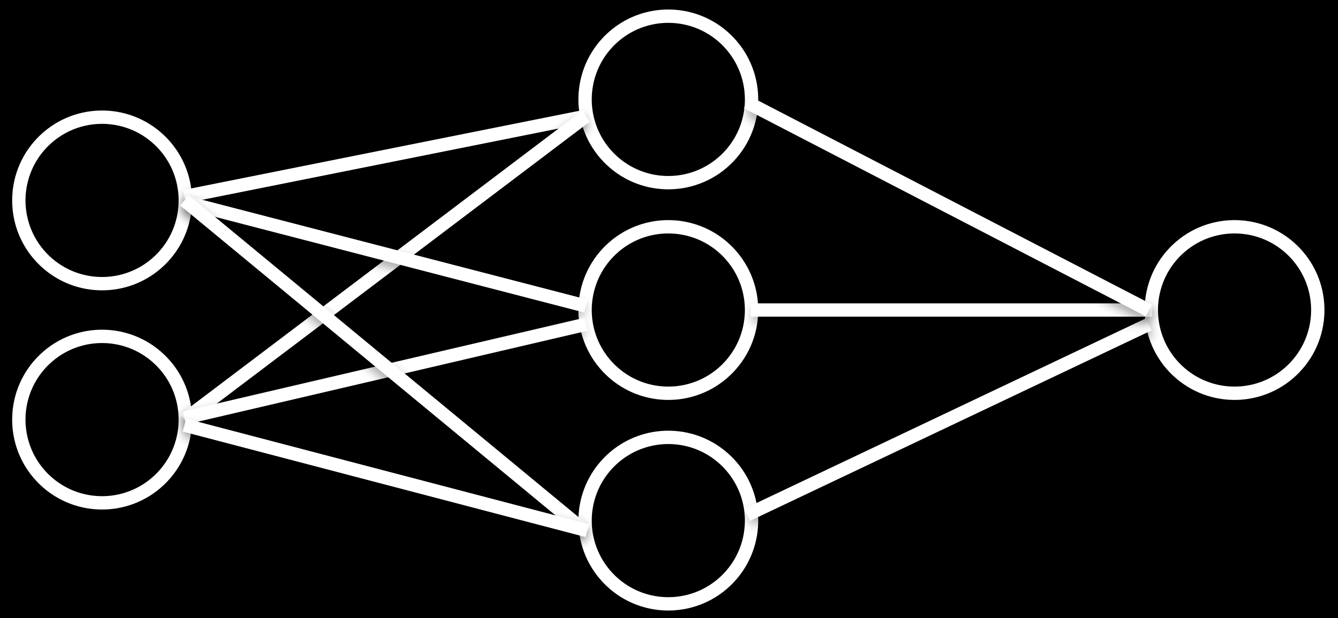


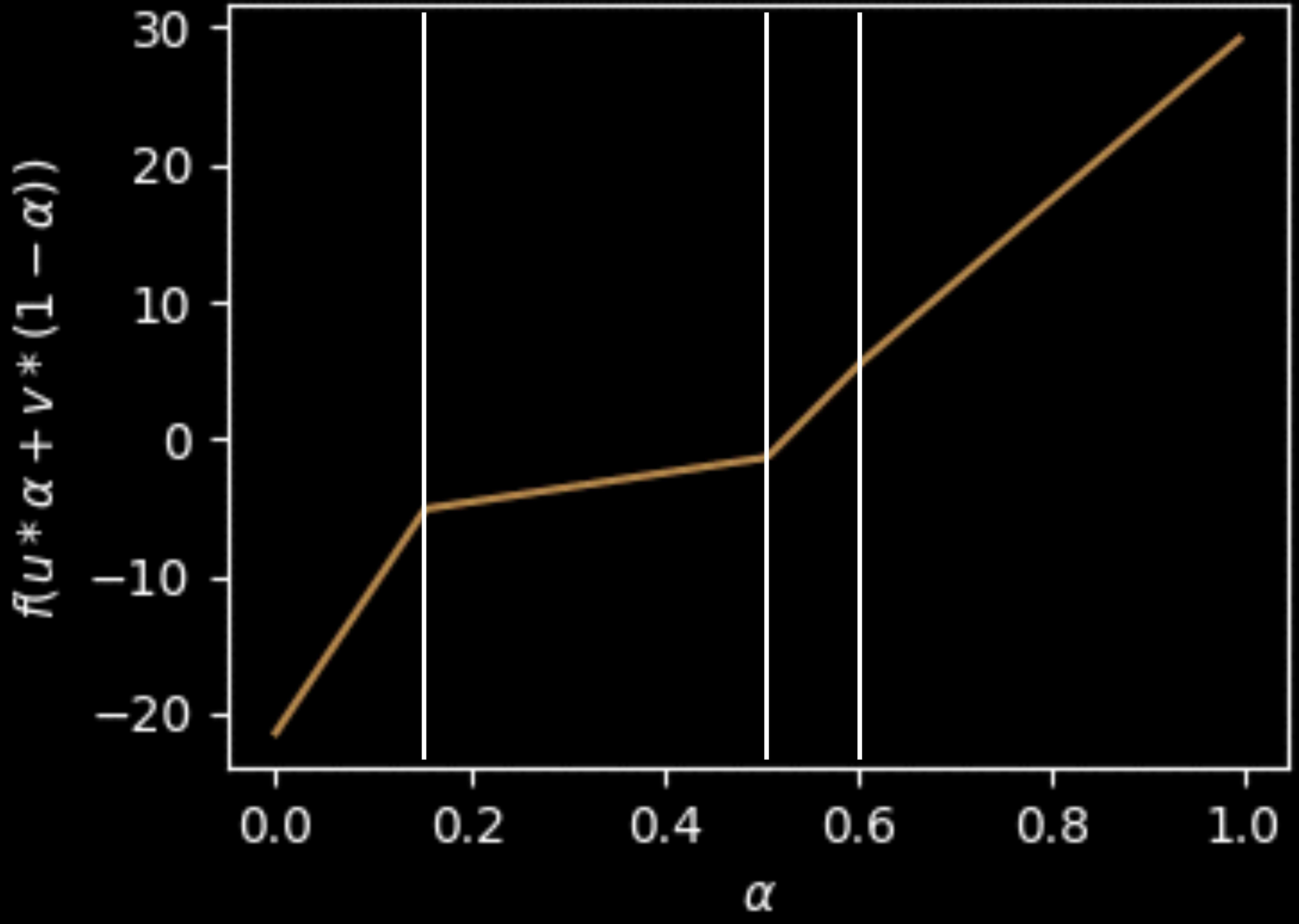


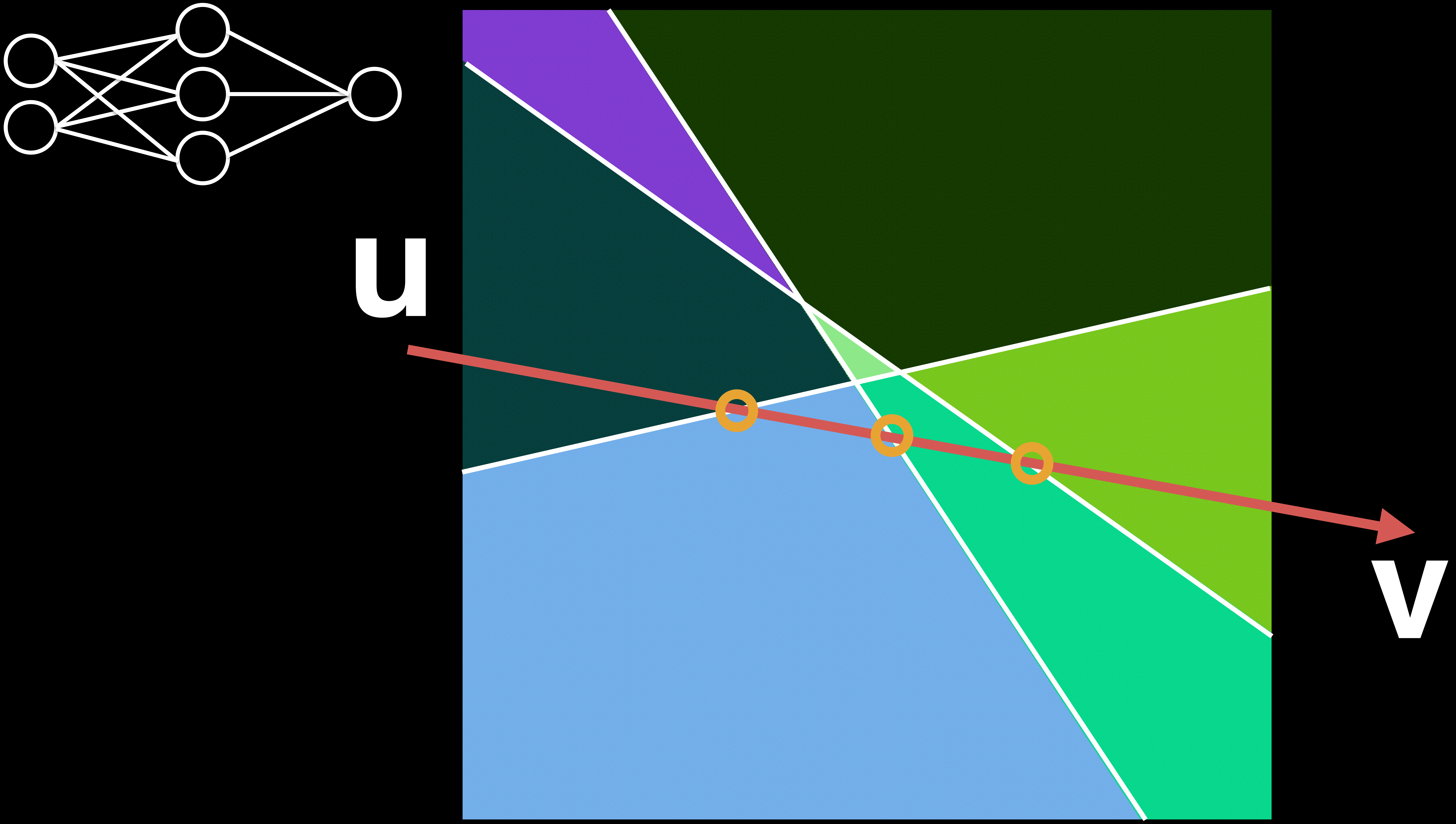


Finding *witnesses* to each neuron


u
v




$f(u*\alpha+v*(1-\alpha))$
30
20
10
0
−10
−20
0.0
0.2
0.4
0.6
0.8
1.0
α


u
v

Our Contributions

Our Contributions

1. Extract deep models
2. Efficient extraction
3. High Fidelity Extraction

Our Contributions

1. Extract deep models
2. Efficient extraction
3. High Fidelity Extraction

Our Contributions

1. Extract deep models
2. Efficient extraction
3. High Fidelity Extraction

Our Contributions

1. Extract 2-deep models
2. Efficient extraction
3. High Fidelity Extraction

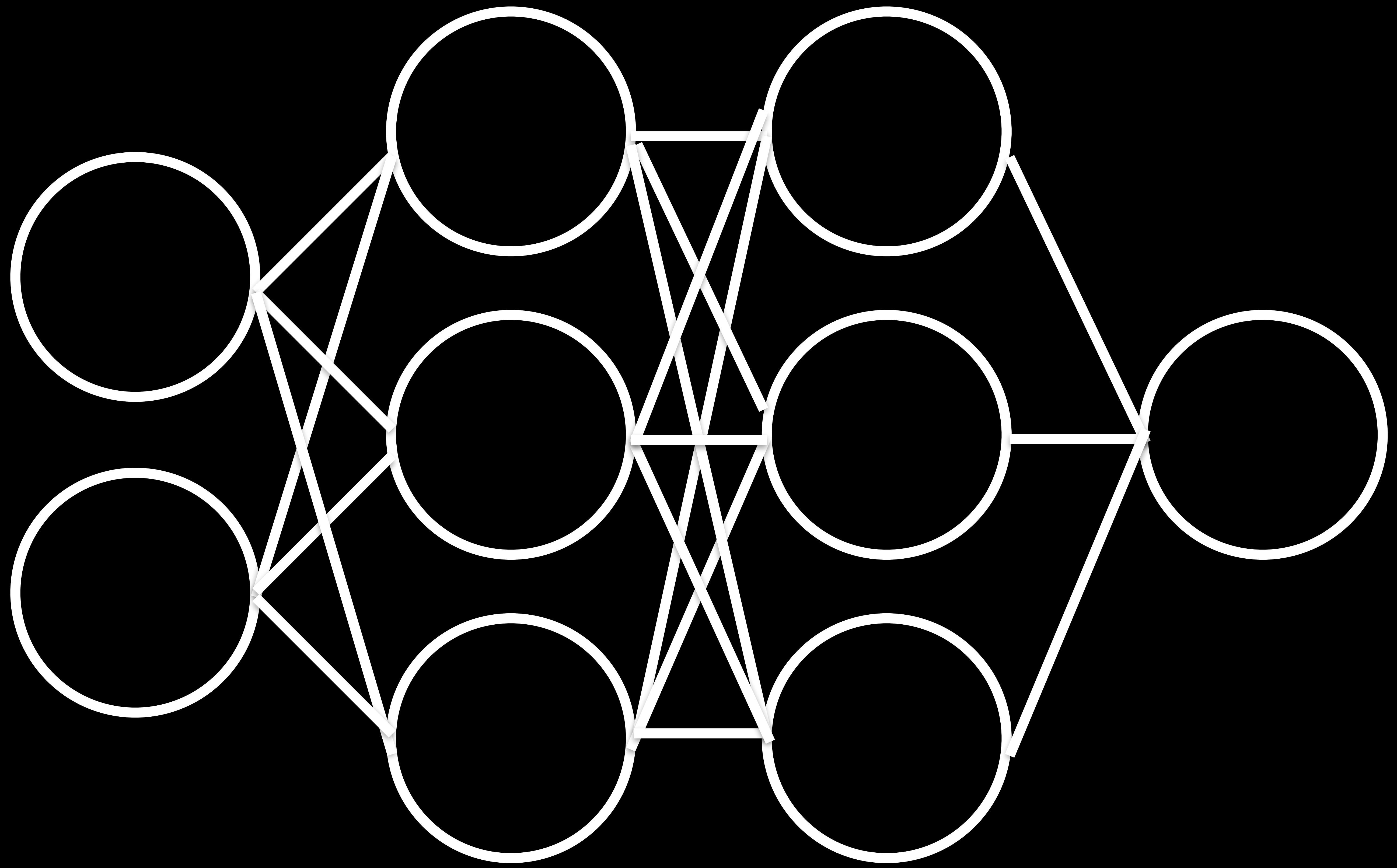
Our Contributions

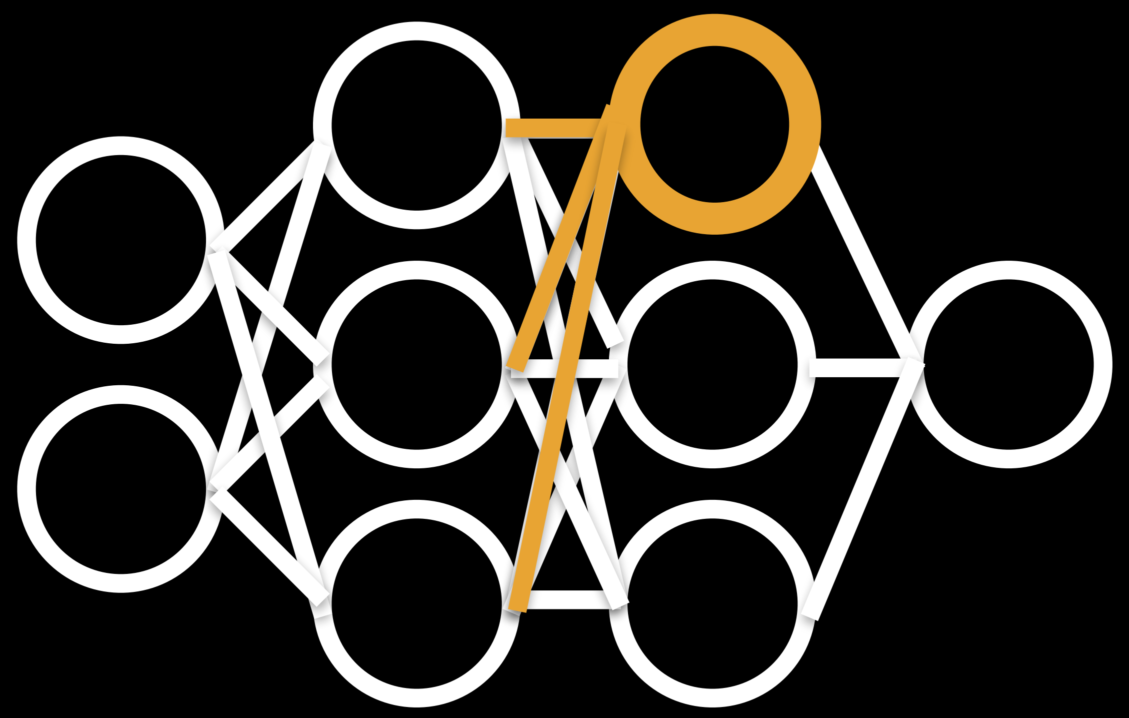
1. Extract 2-deep models
2. Efficient extraction
3. High Fidelity Extraction

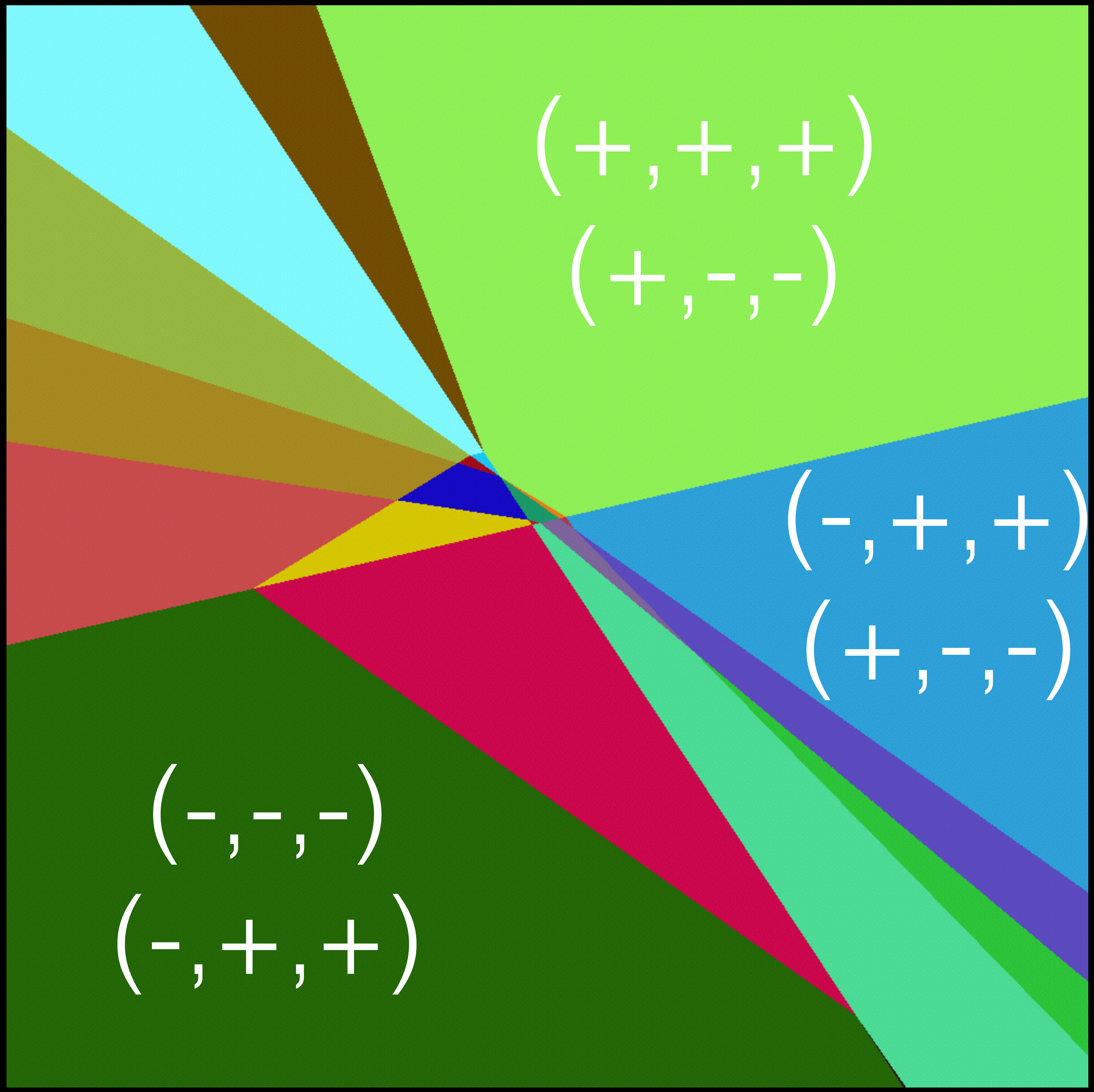
Our Contributions

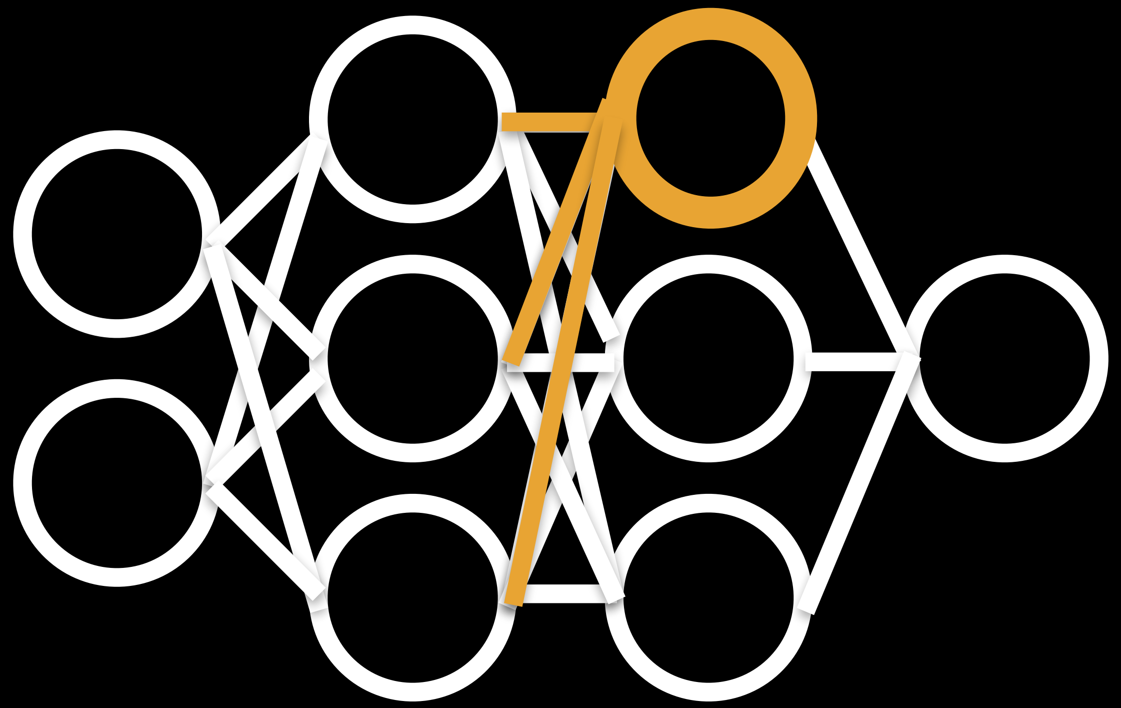
1. Extract 2-deep models
    a. Recover weight values
    b. Recover neuron signs

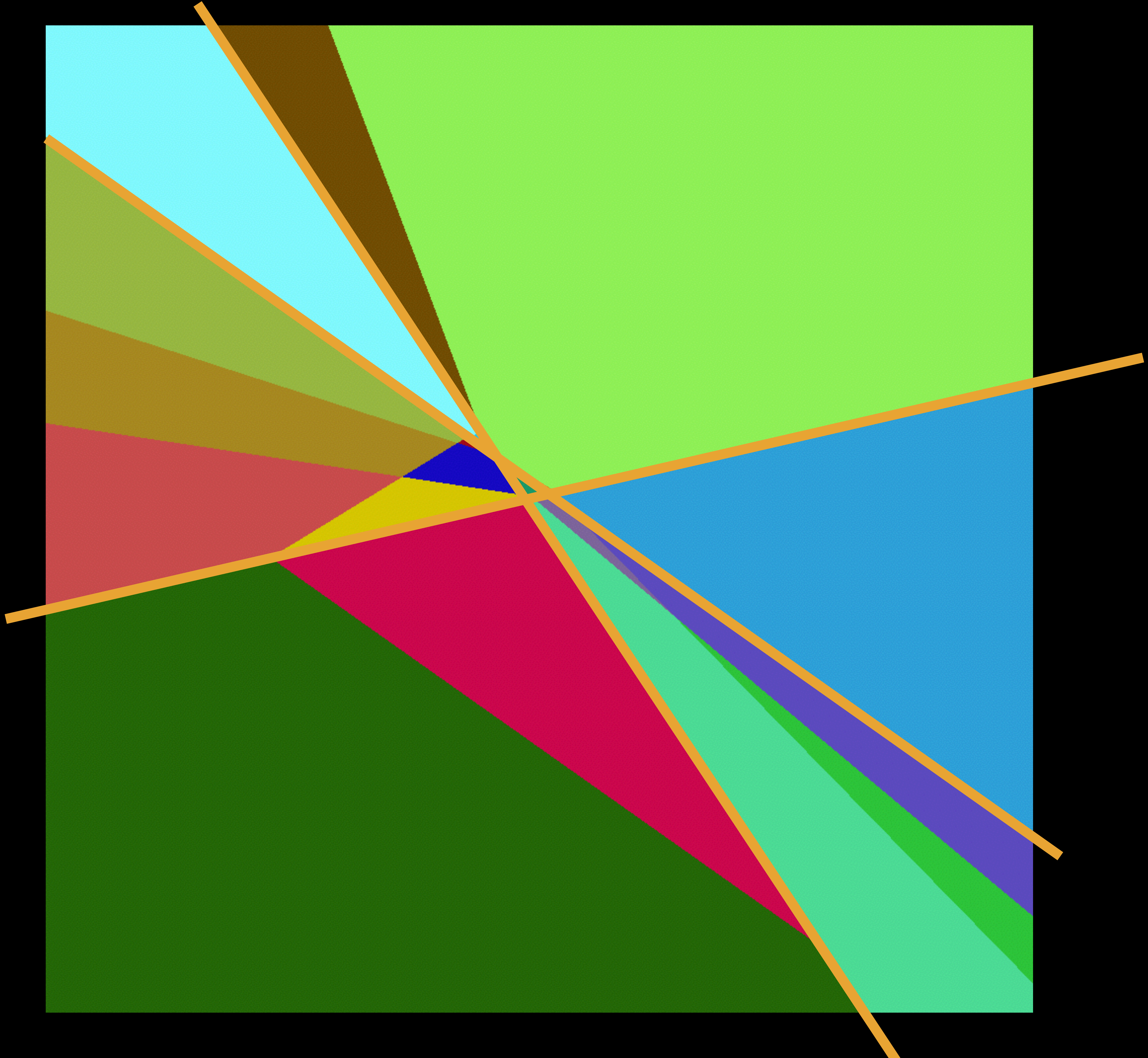
2-deep Neural Network

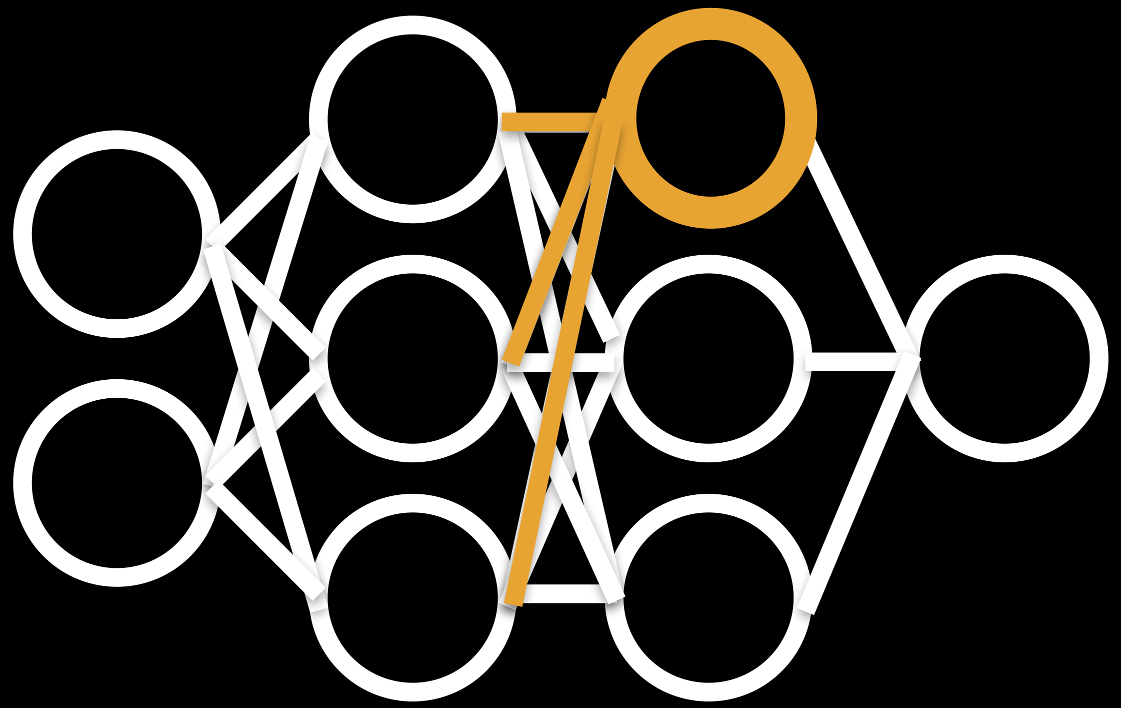


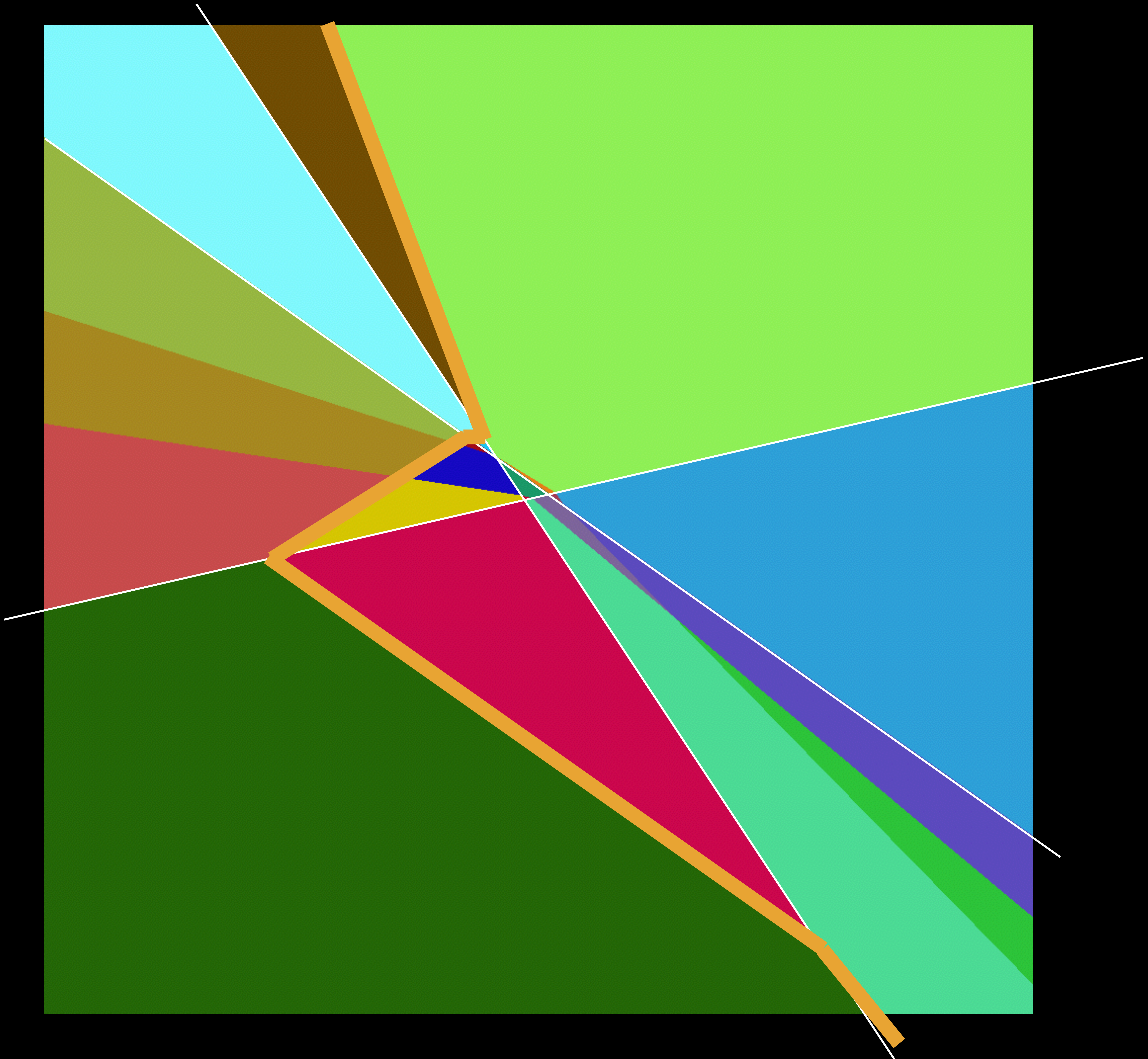






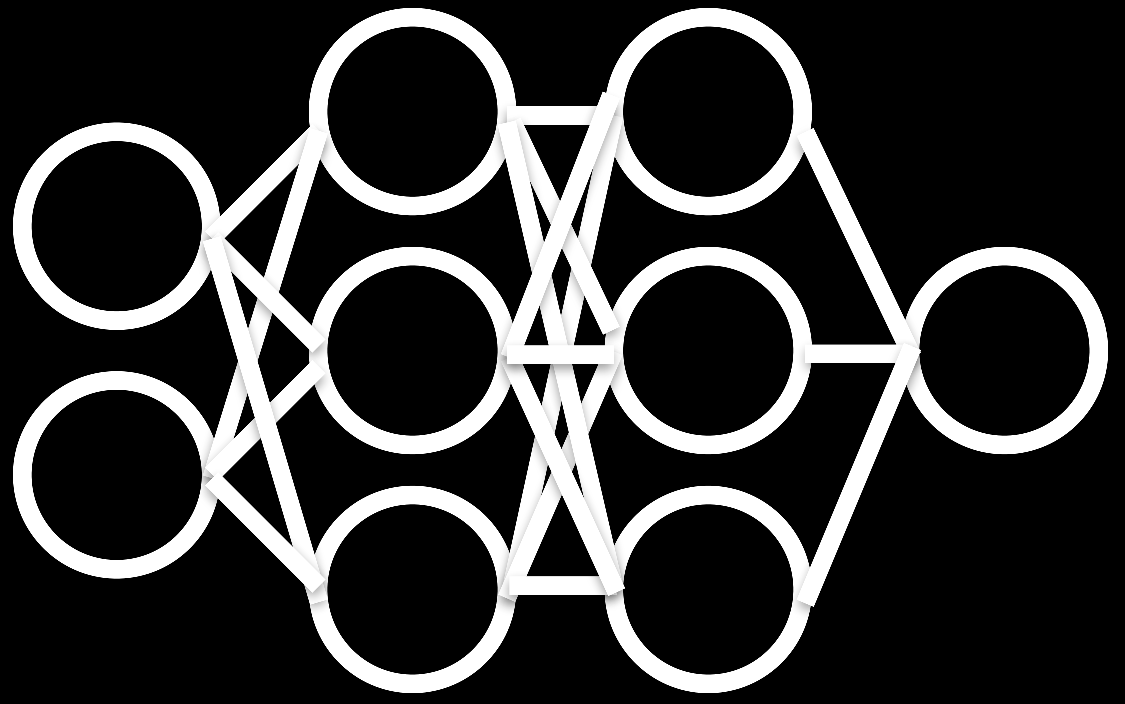


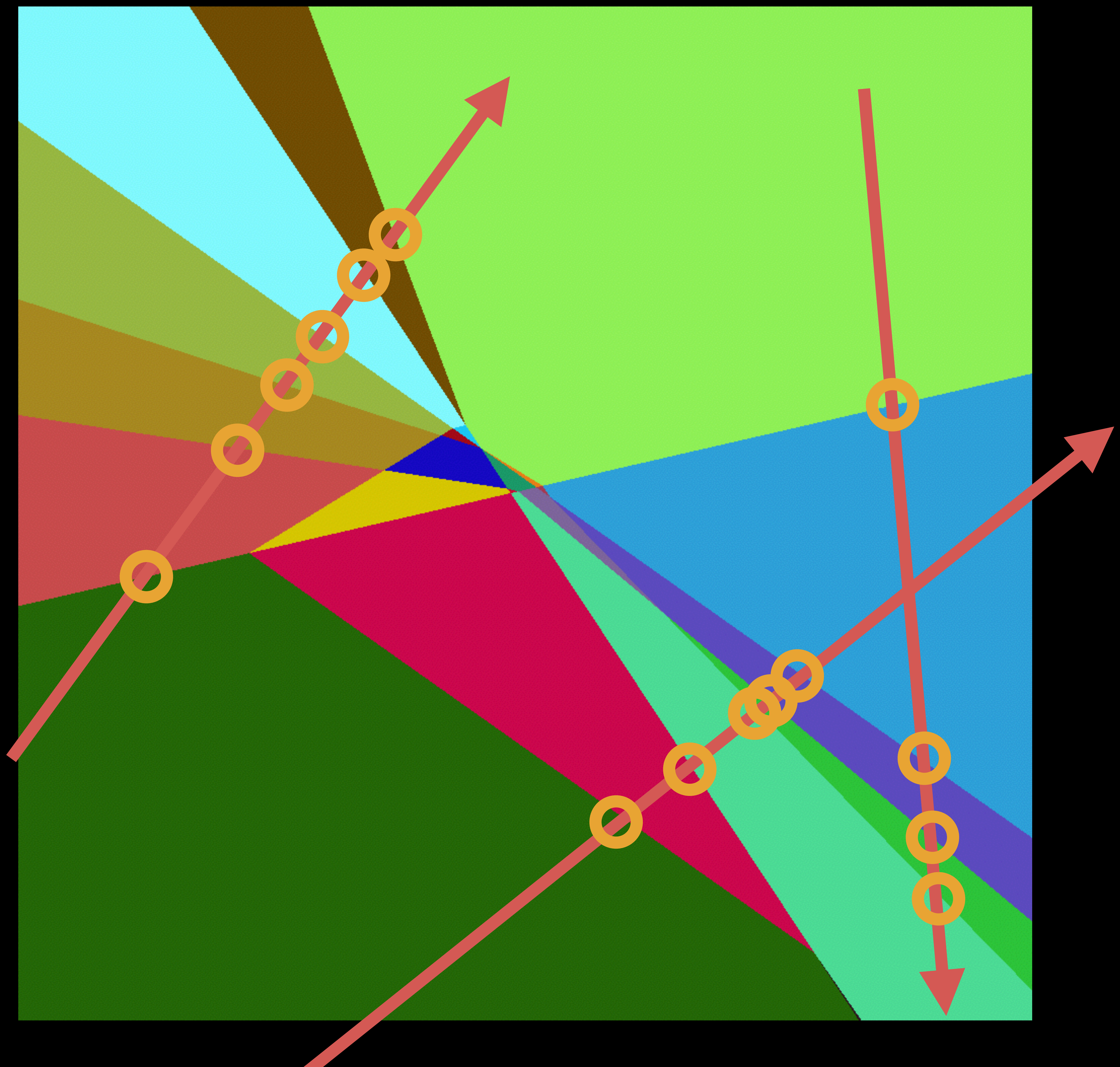


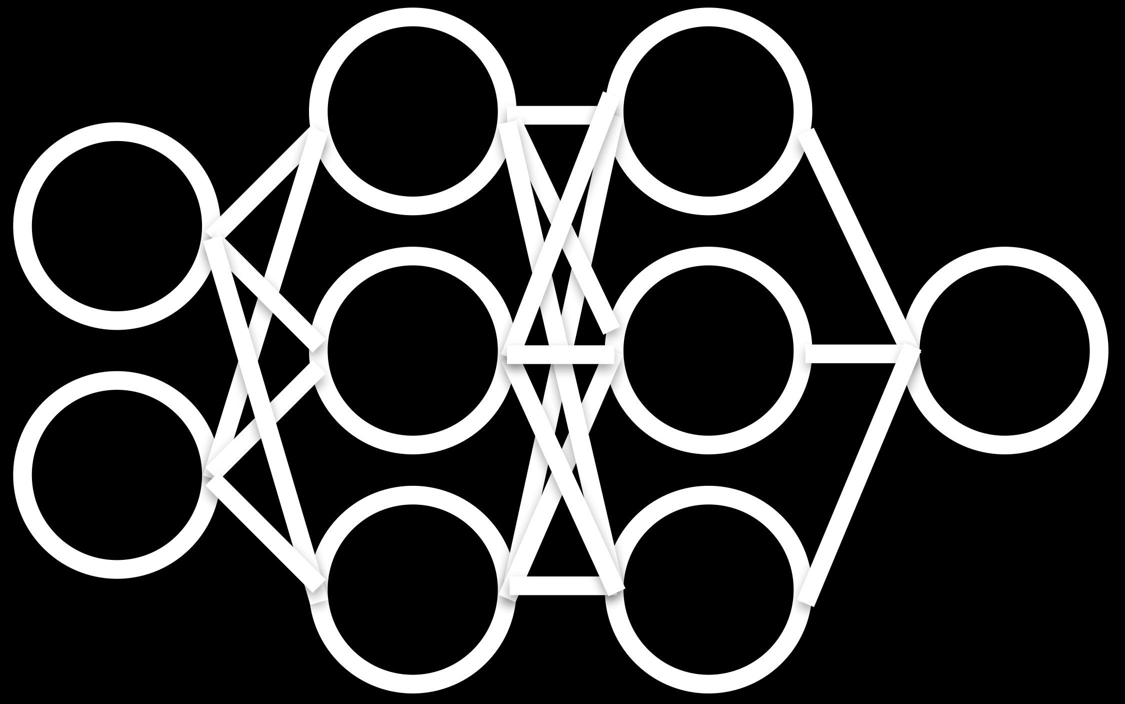


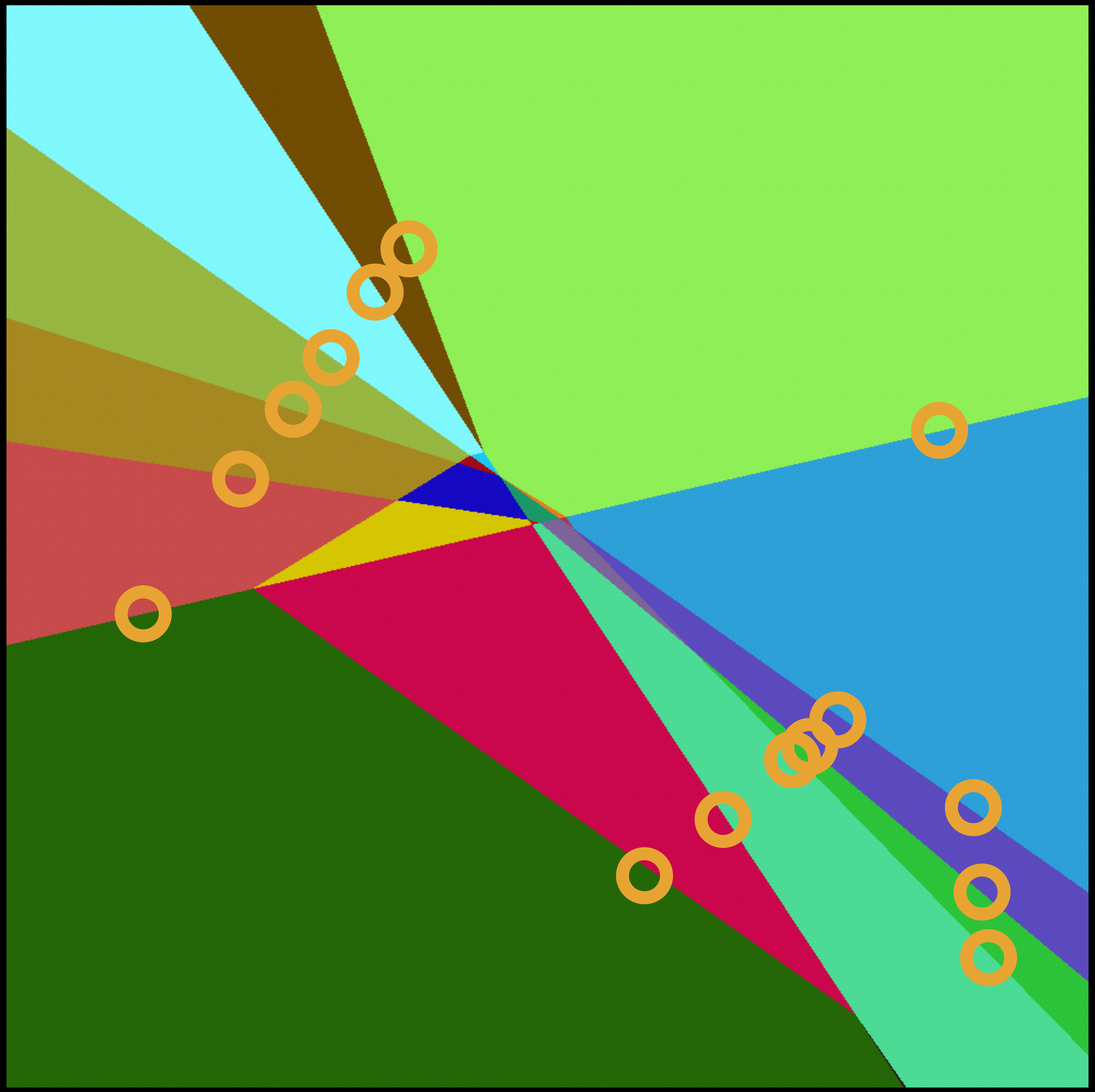
Recovering the
first layer

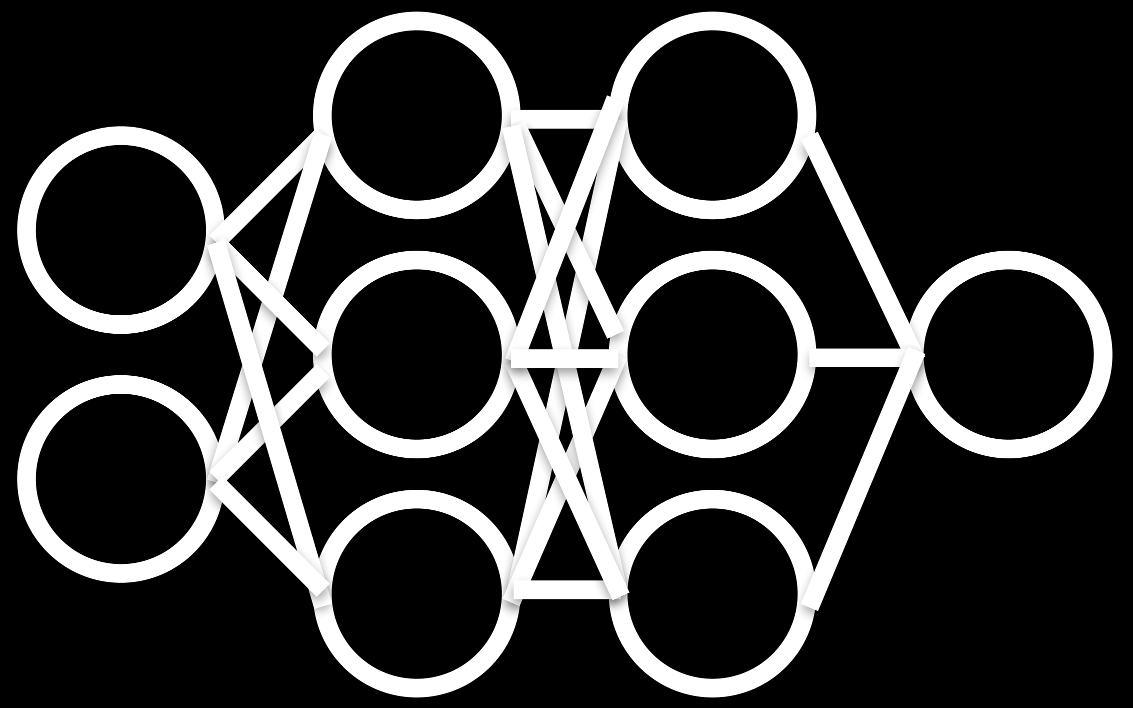
(up to sign)

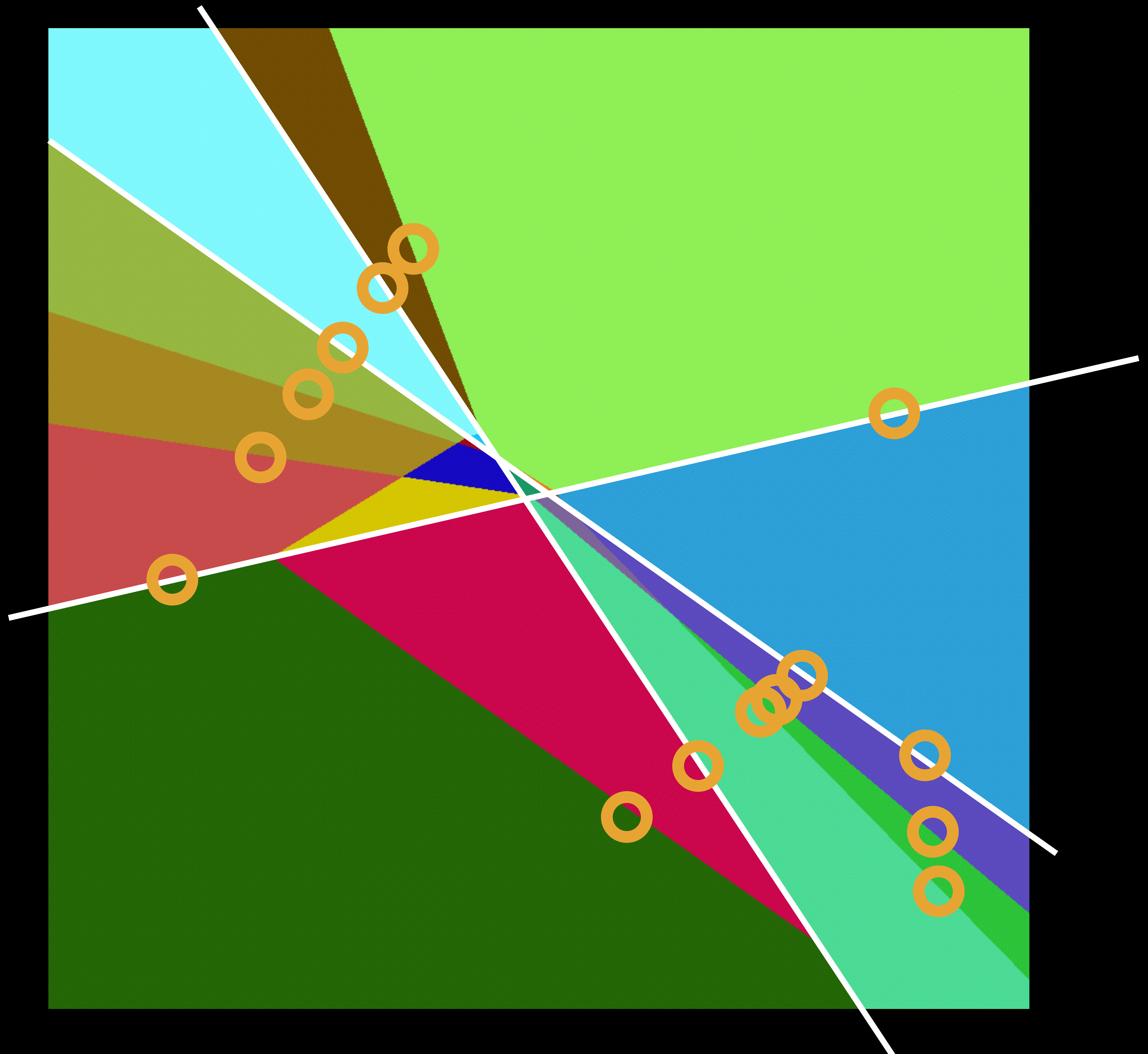


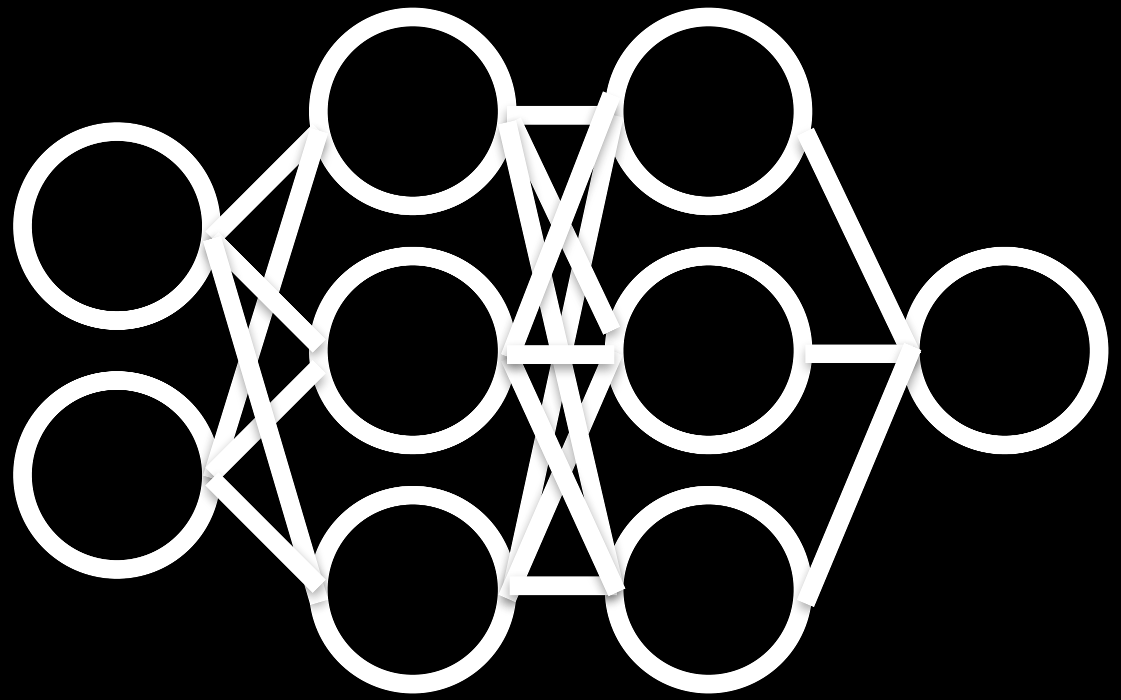


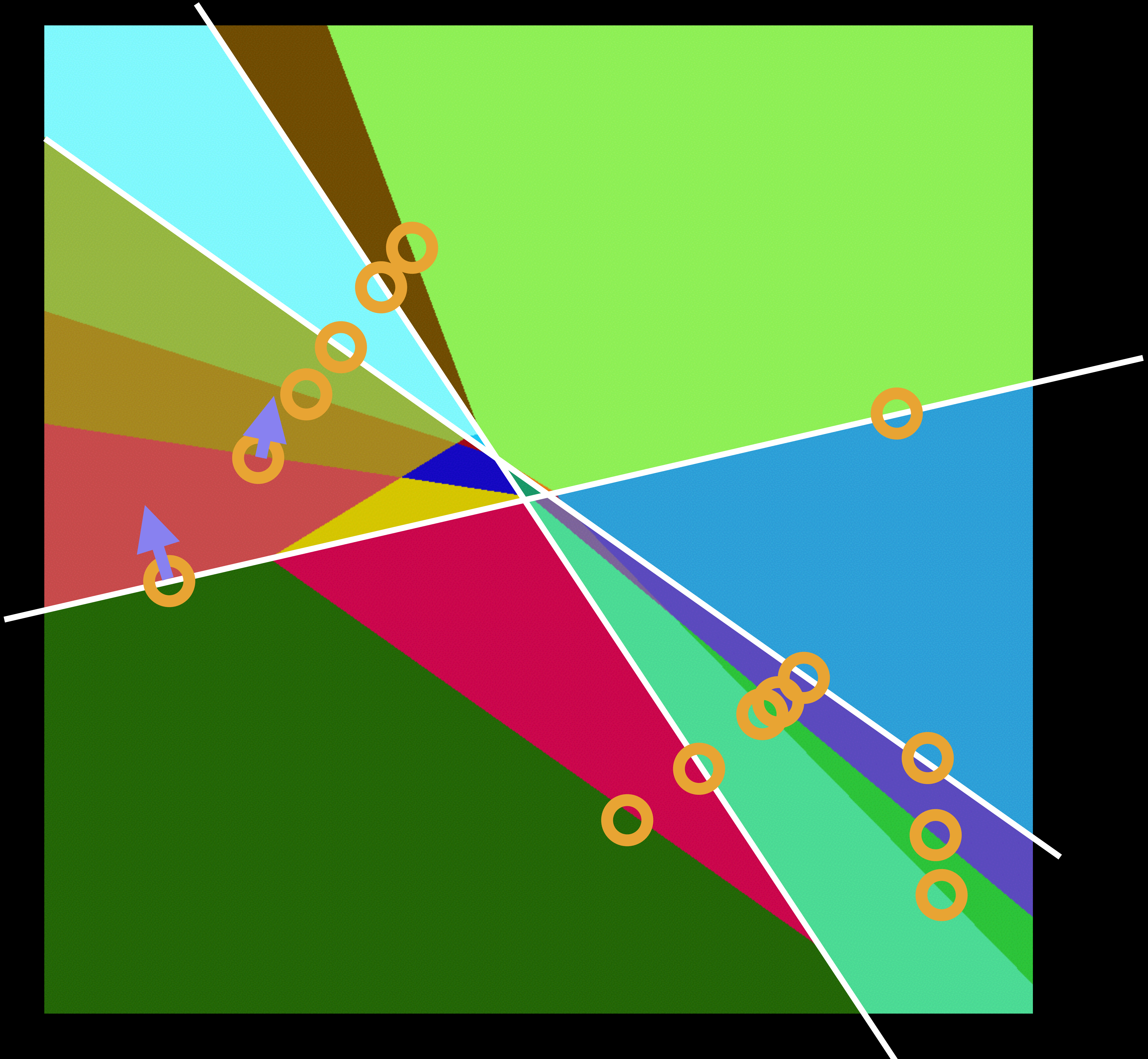


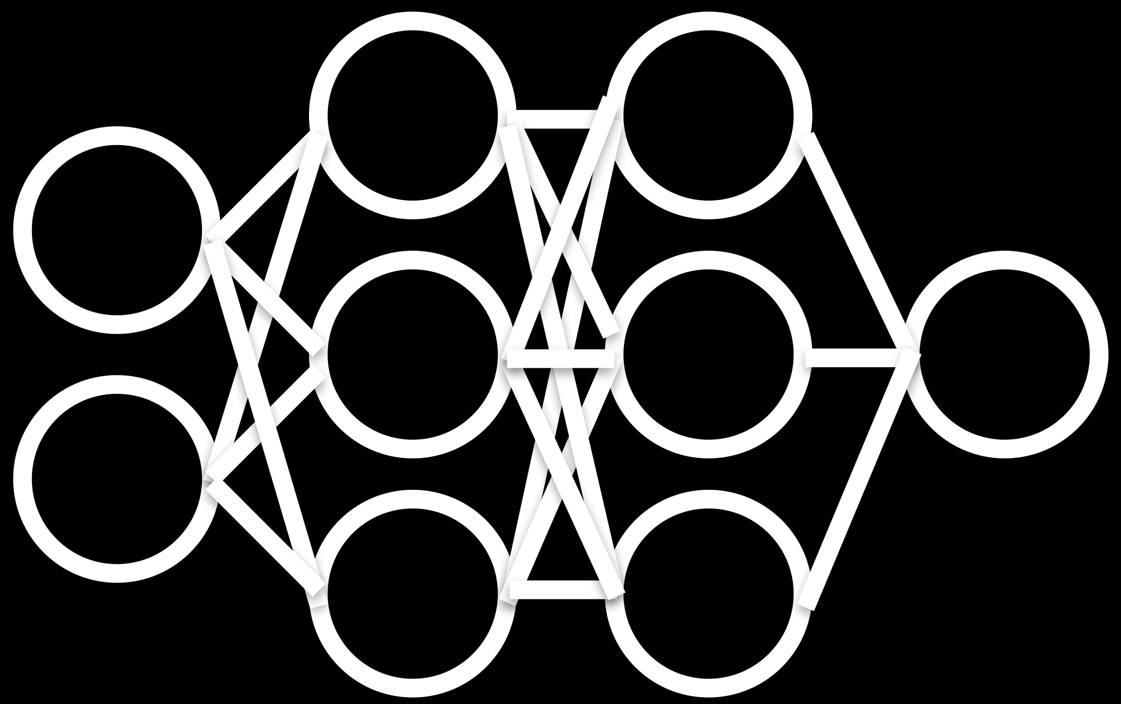


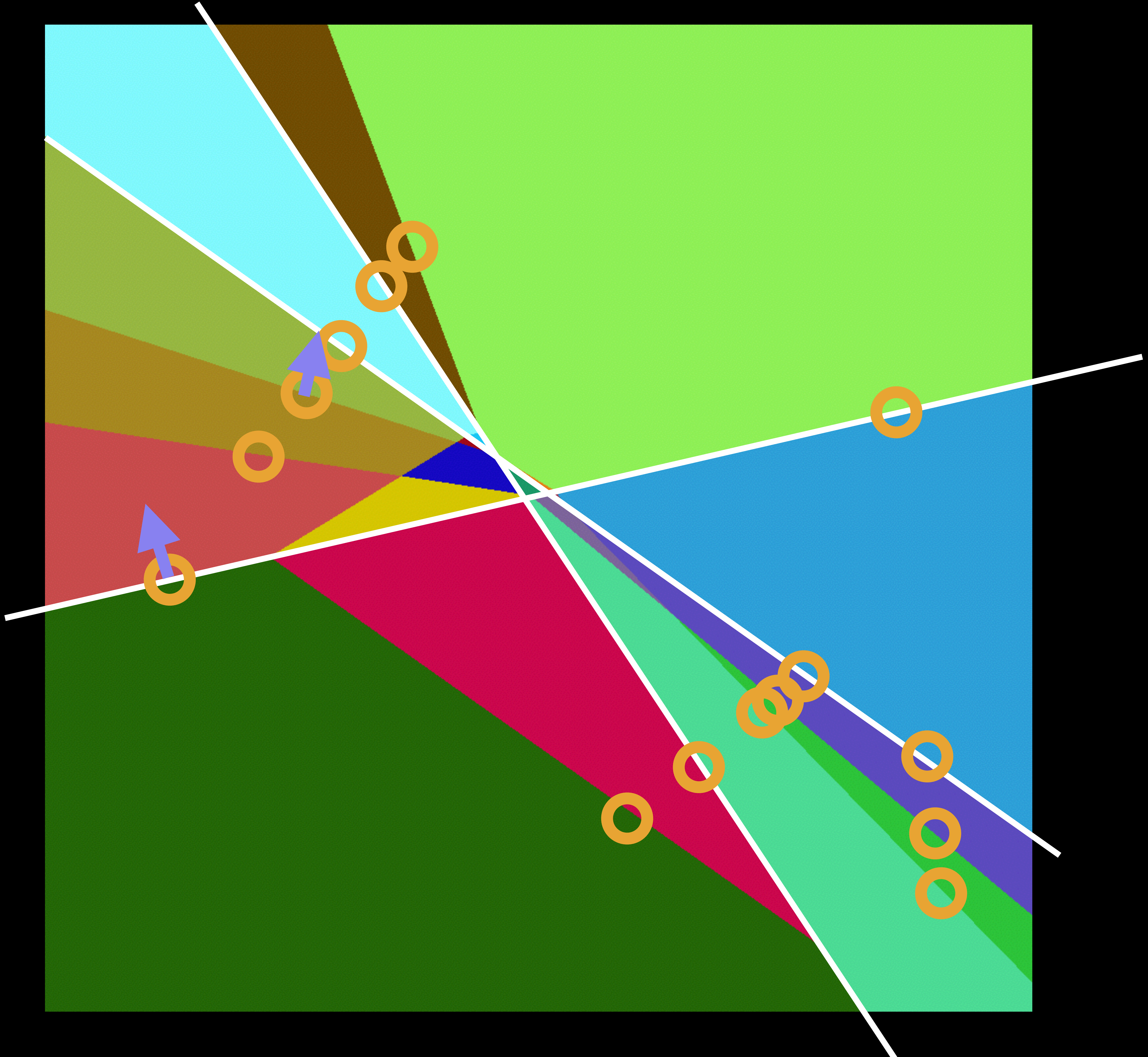


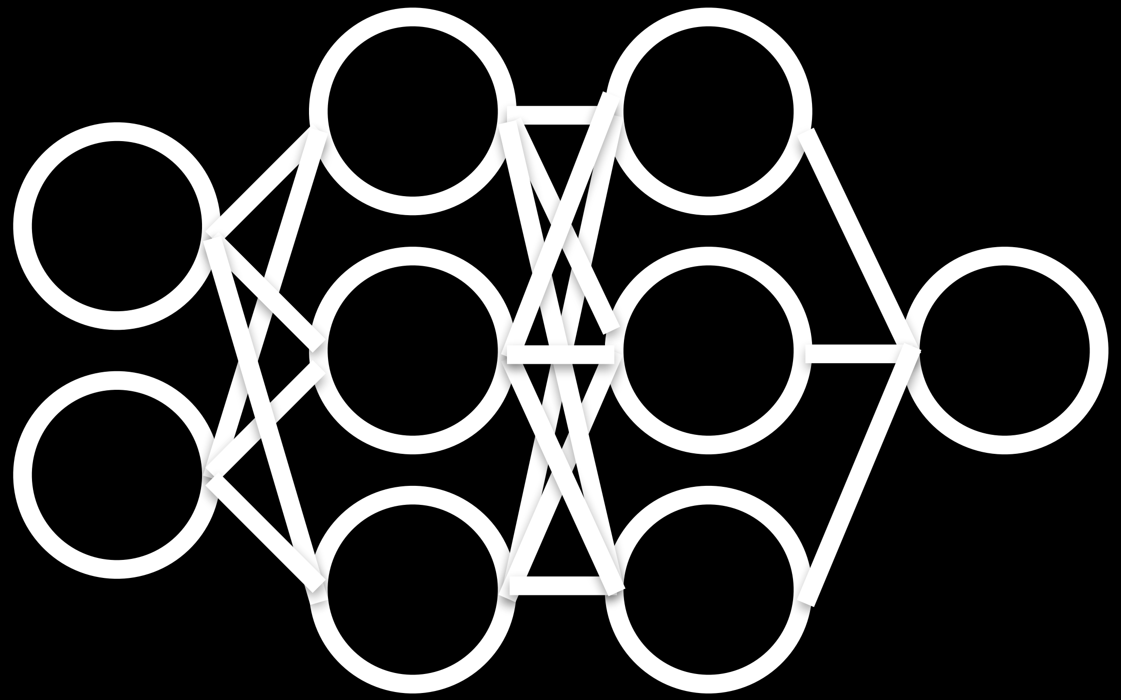


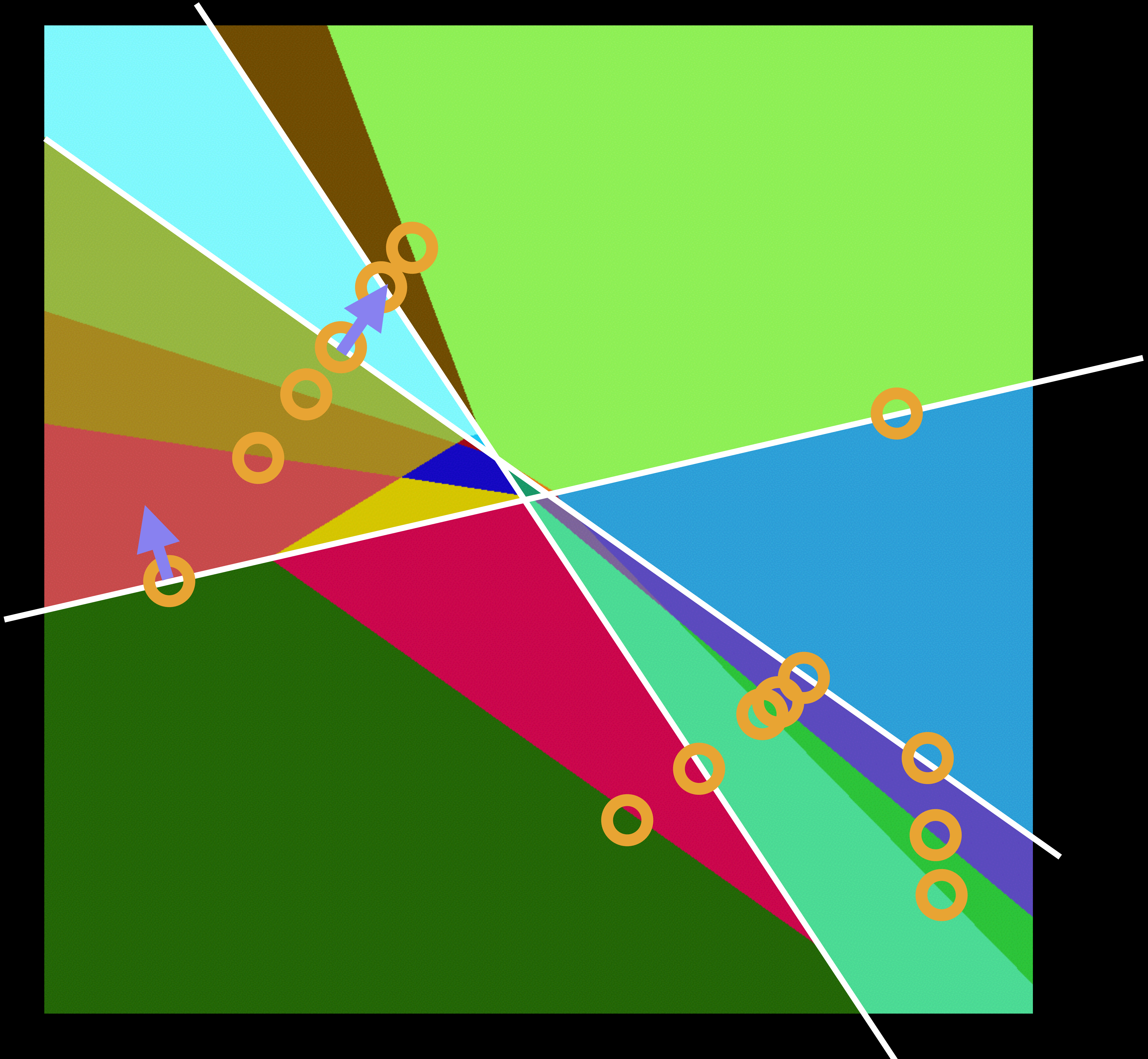


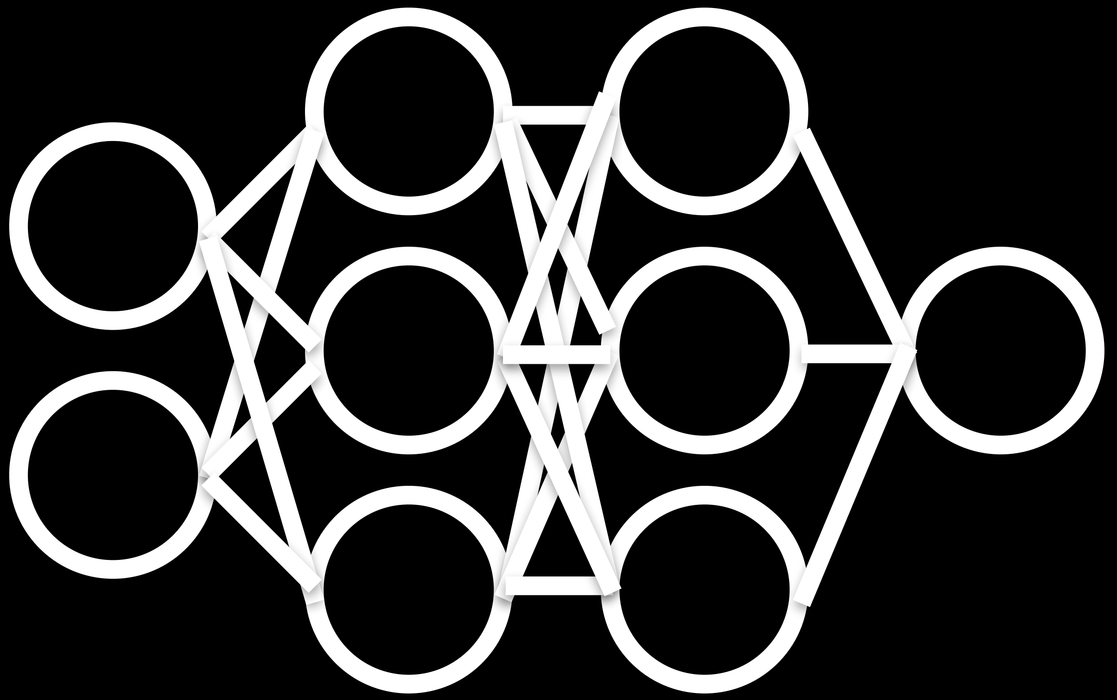


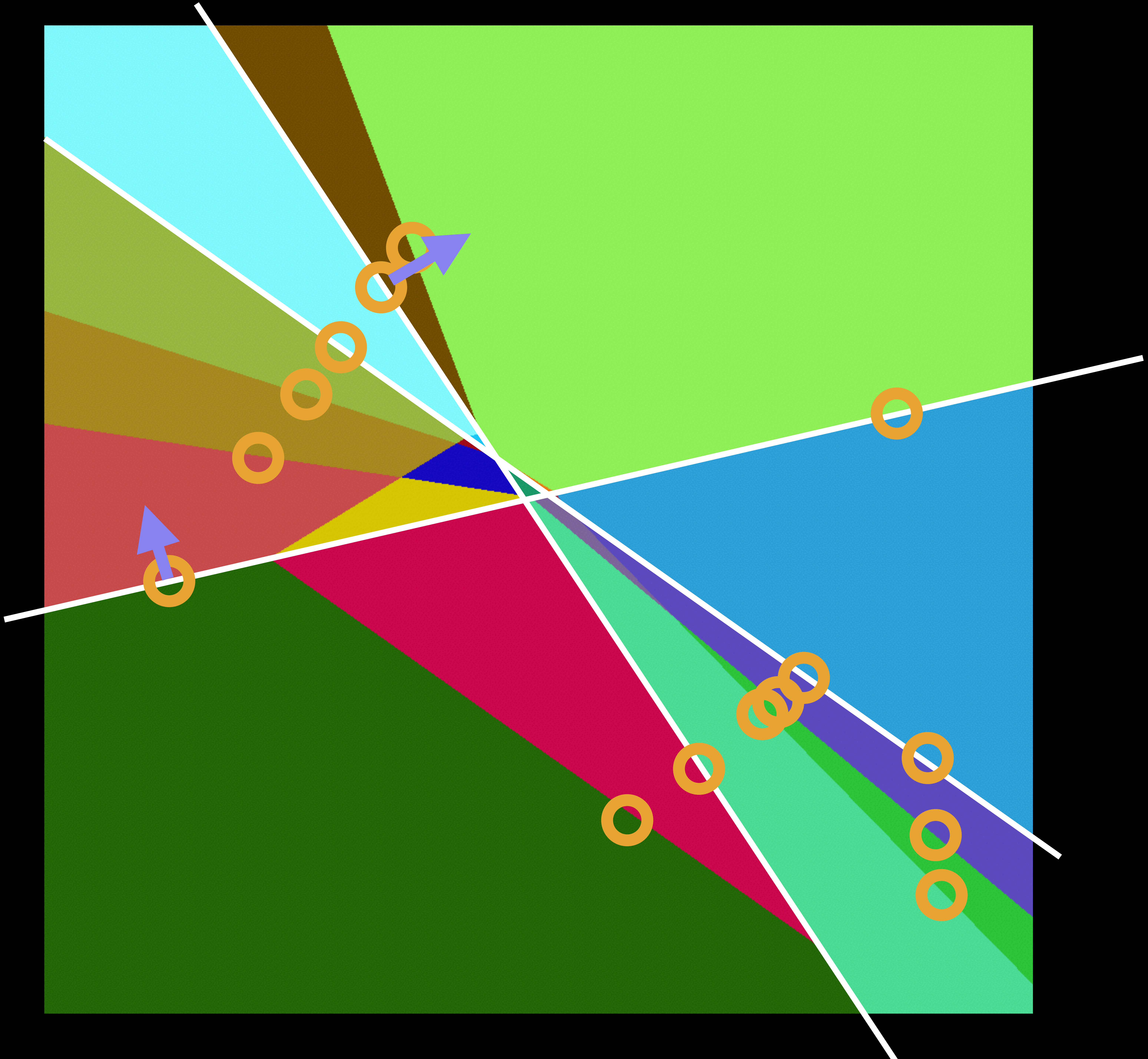


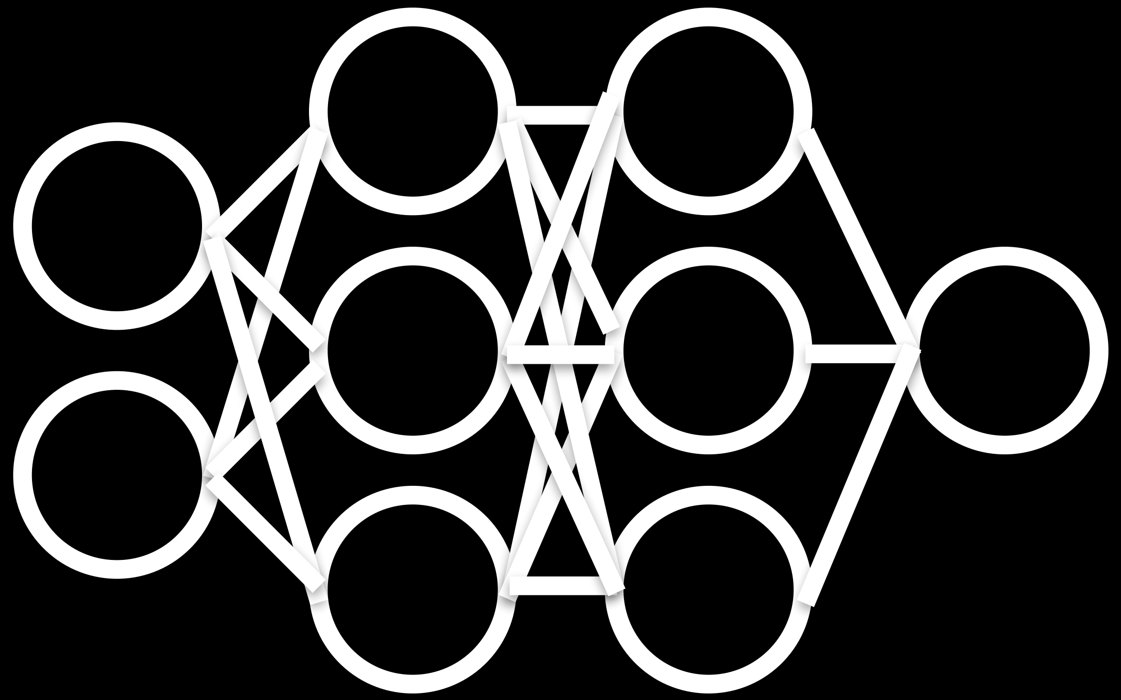


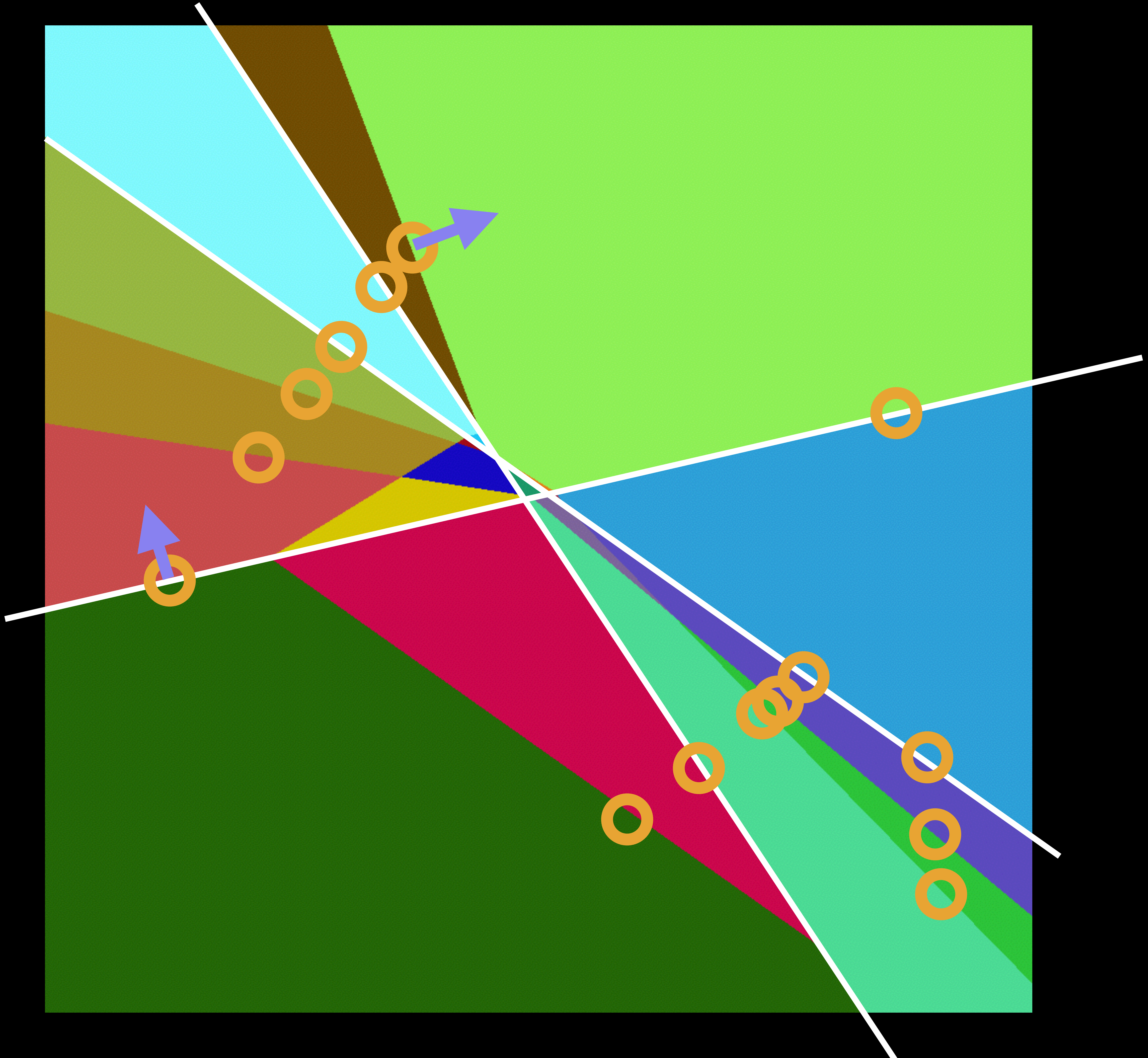


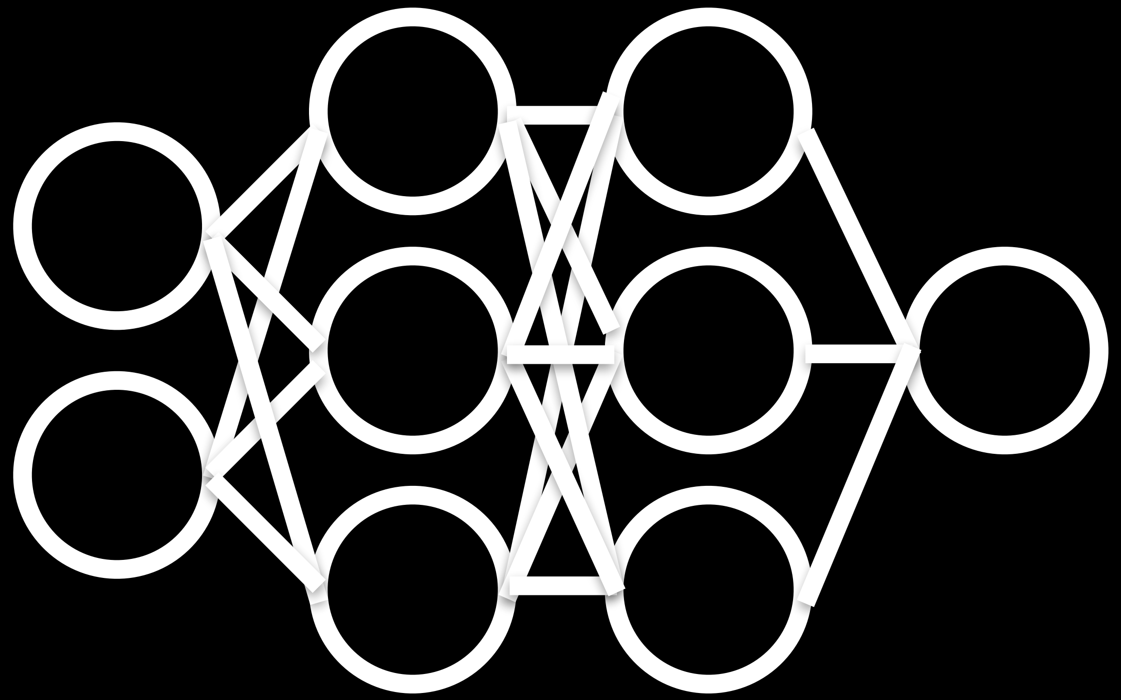


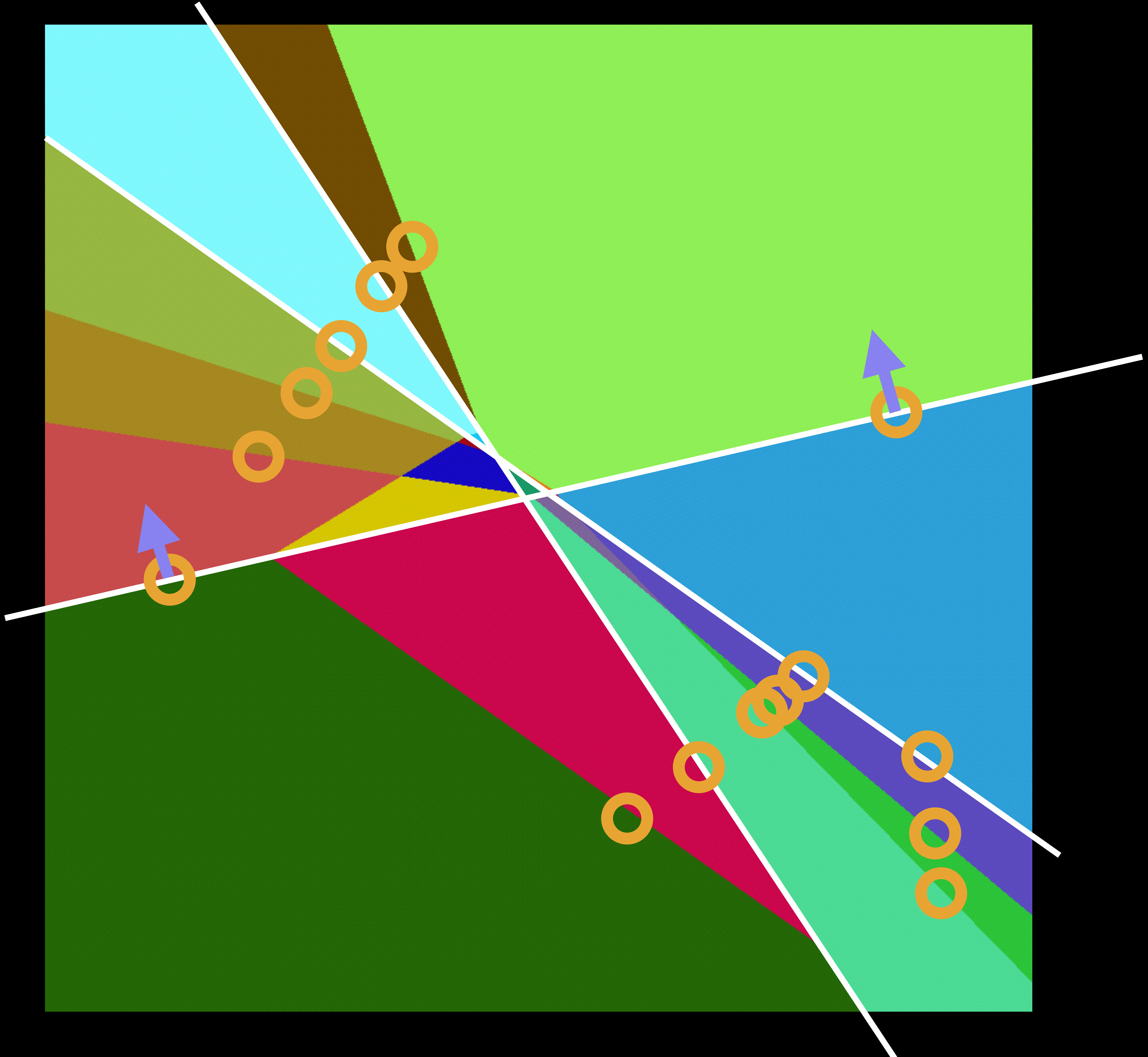


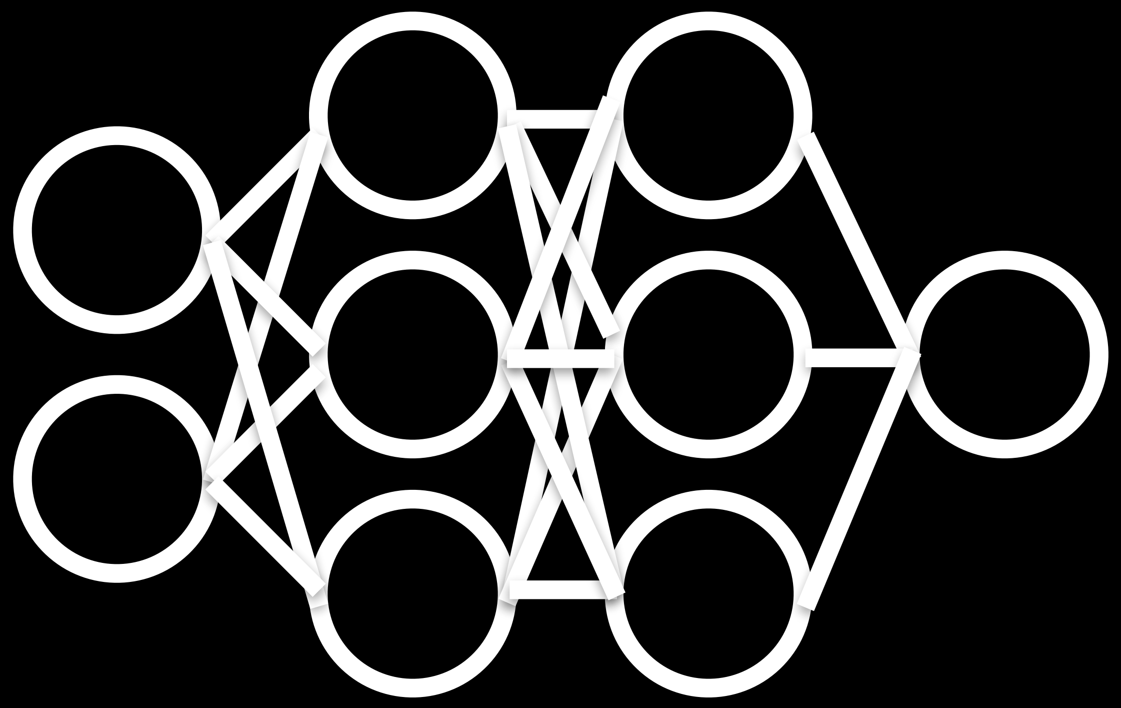


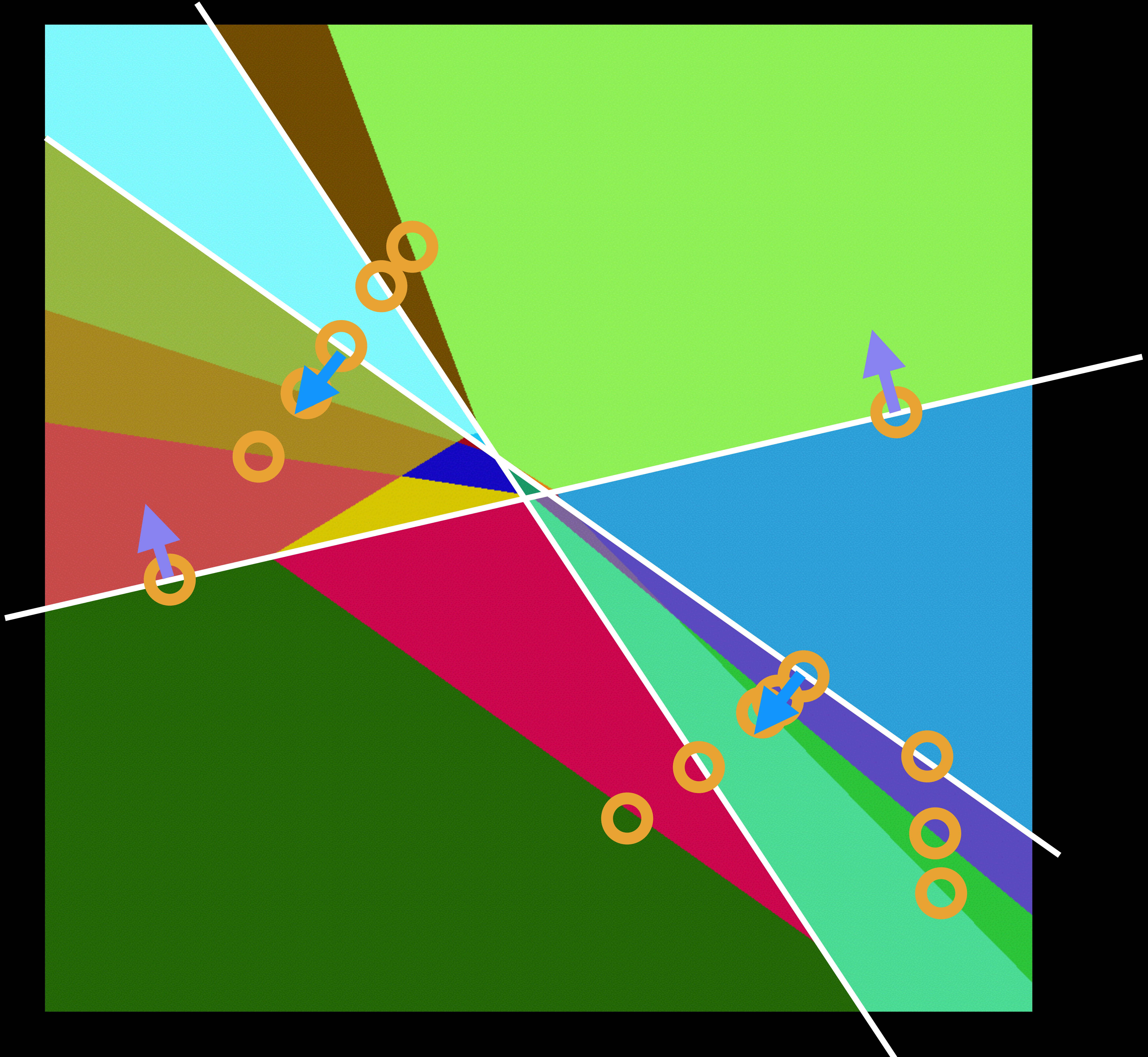


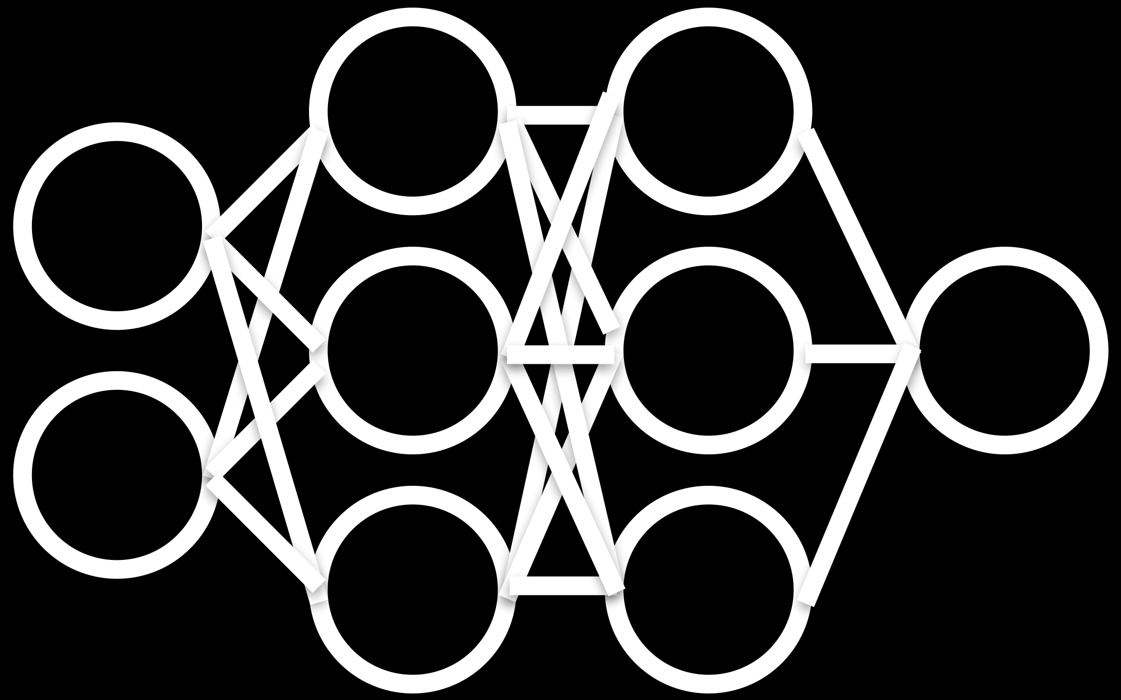


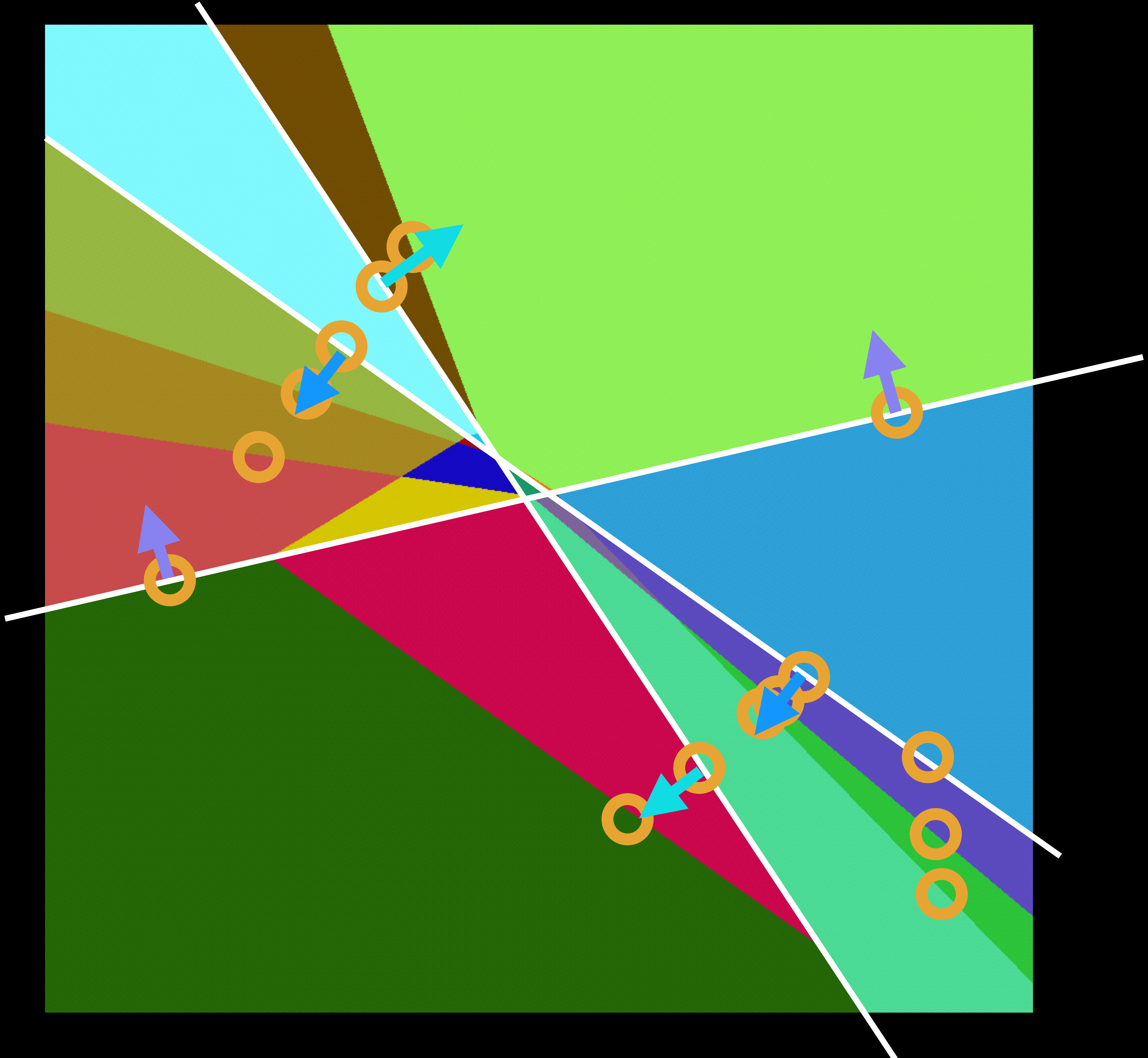


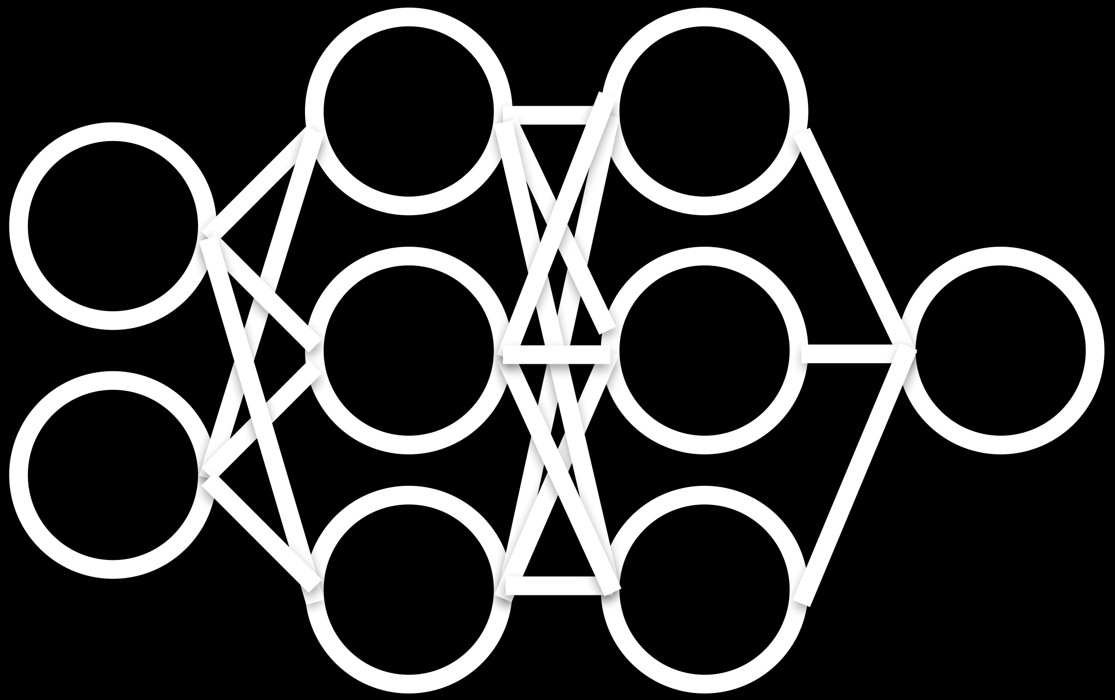


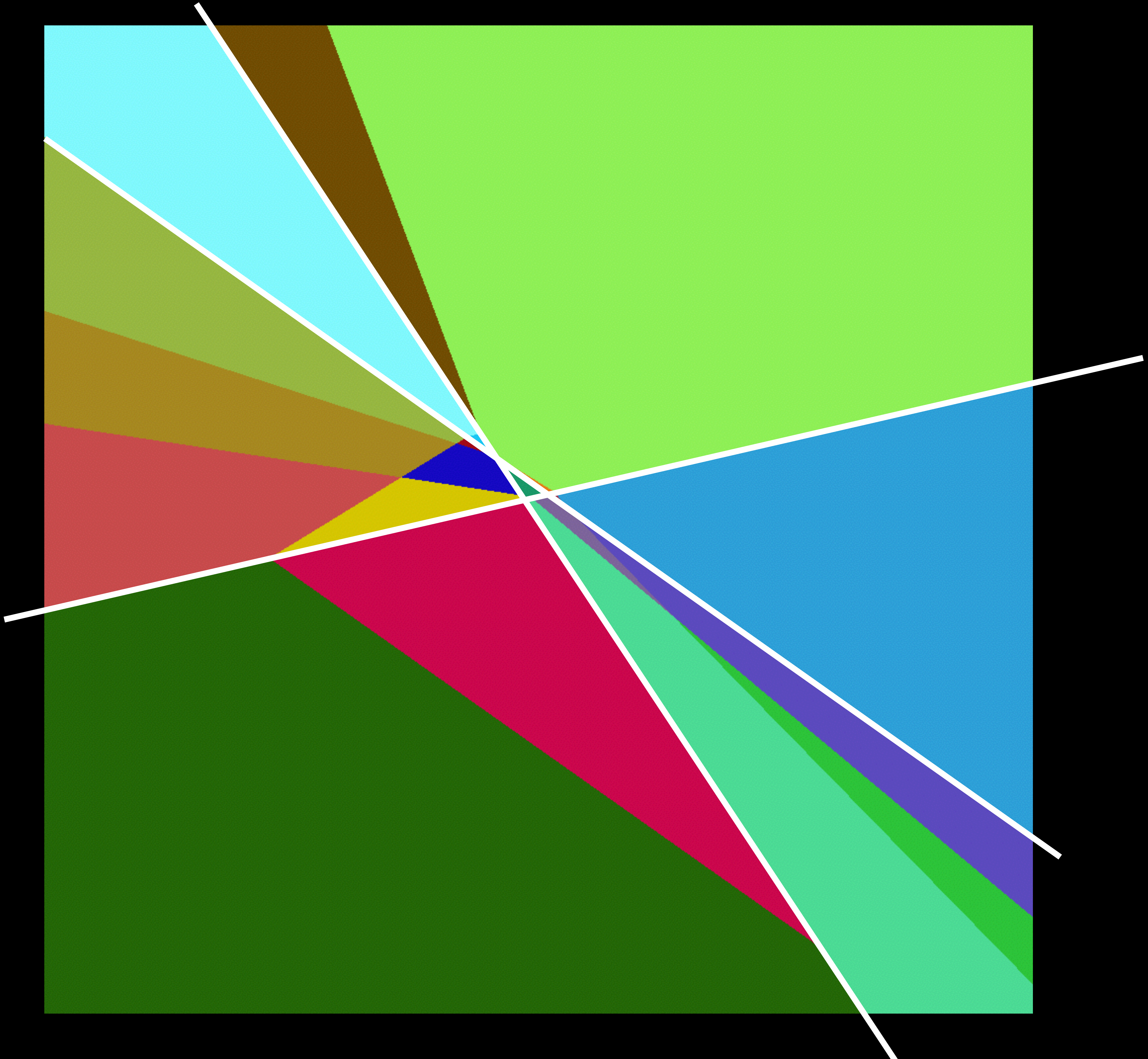


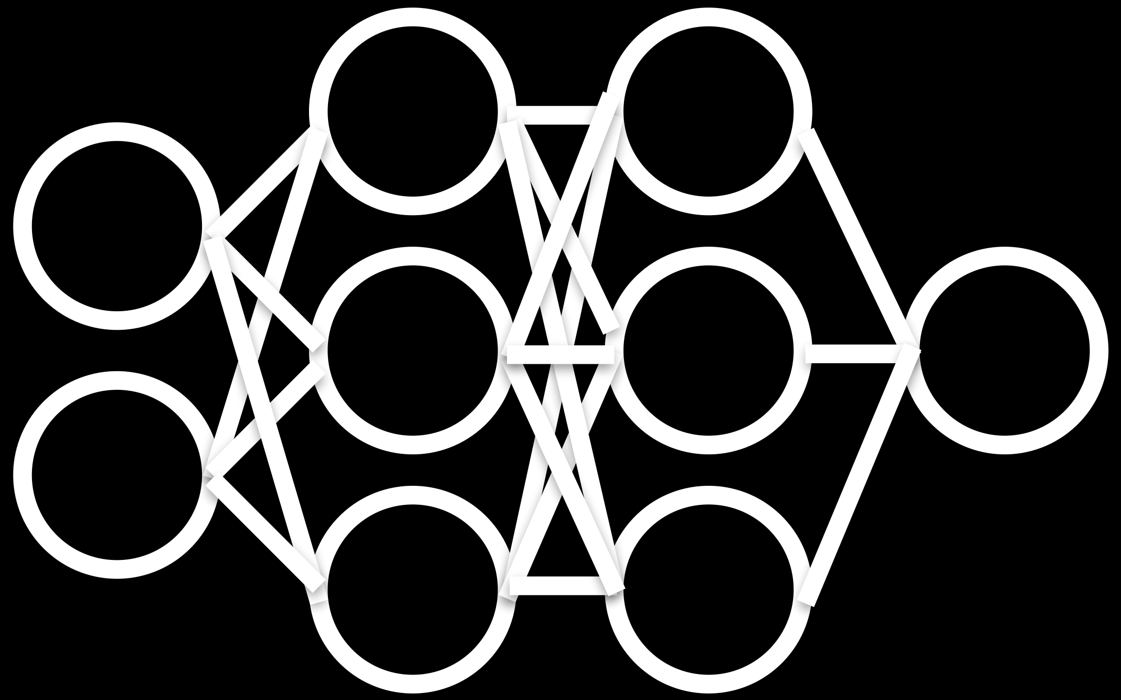


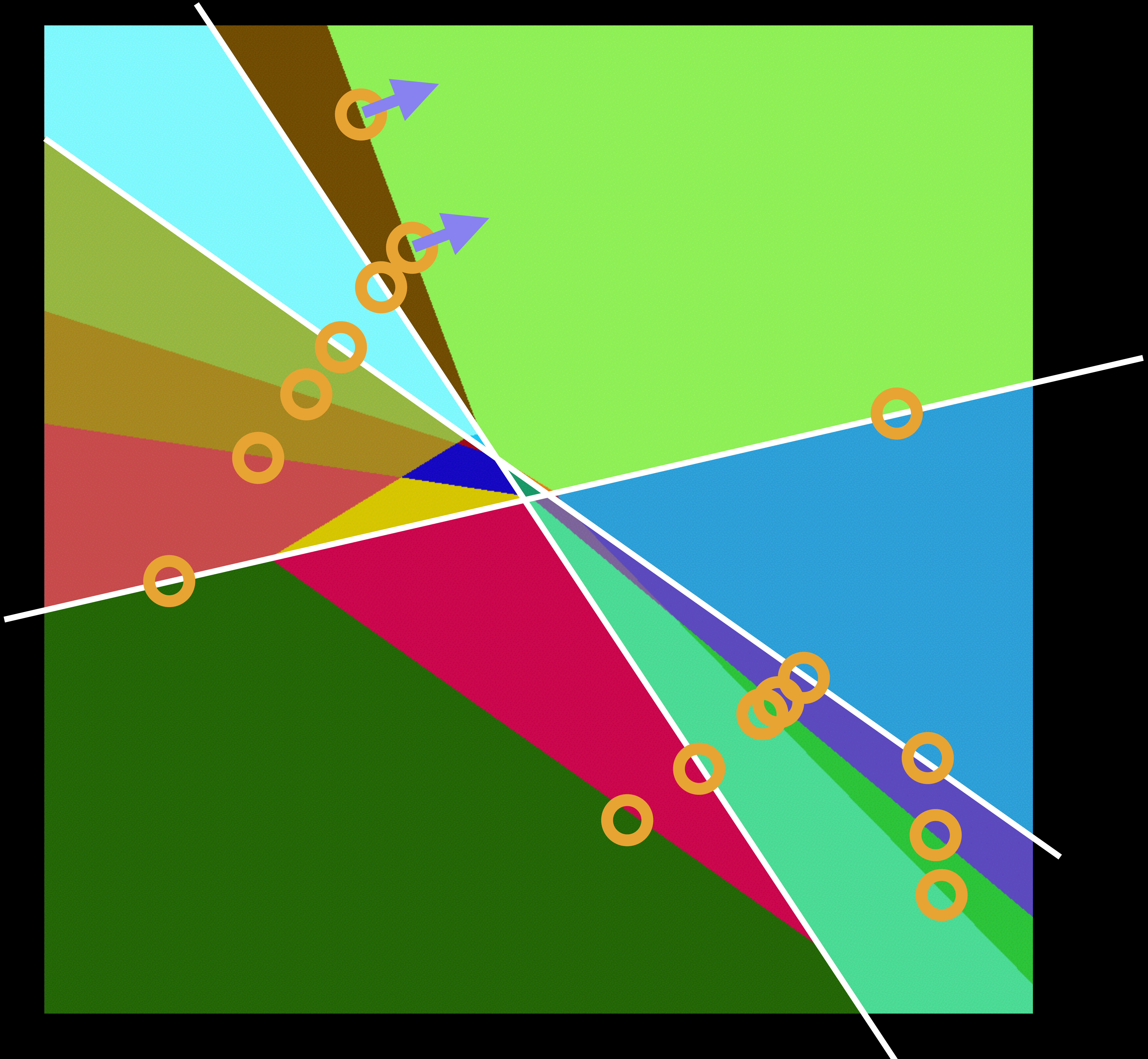






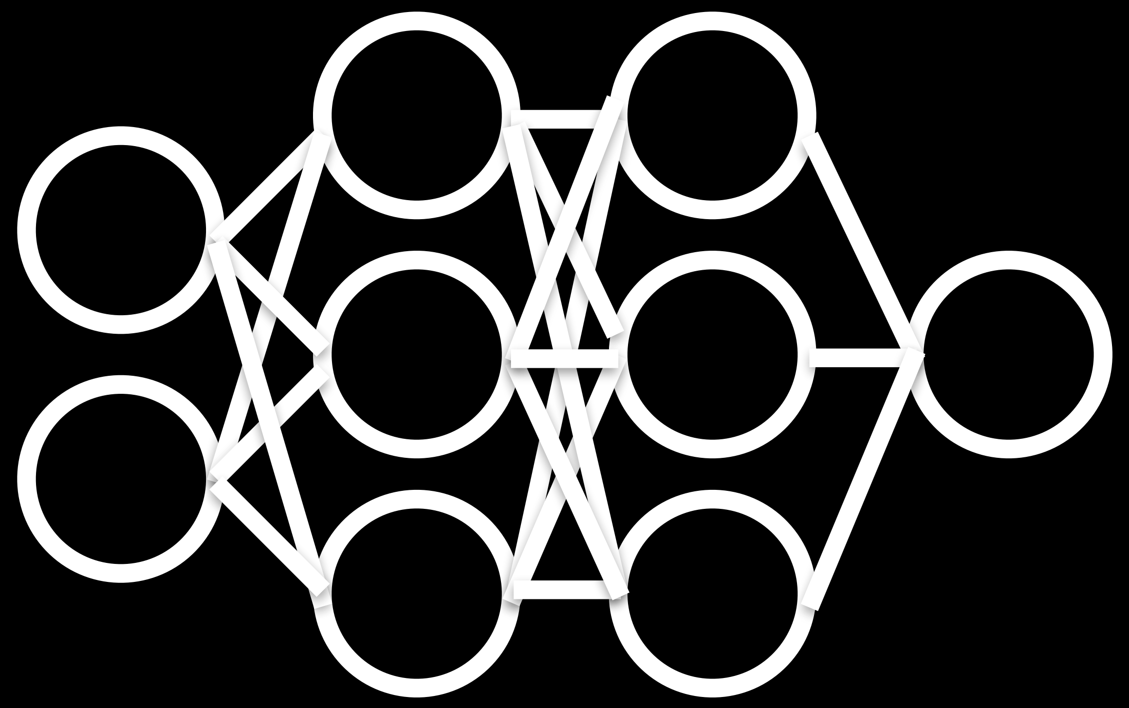


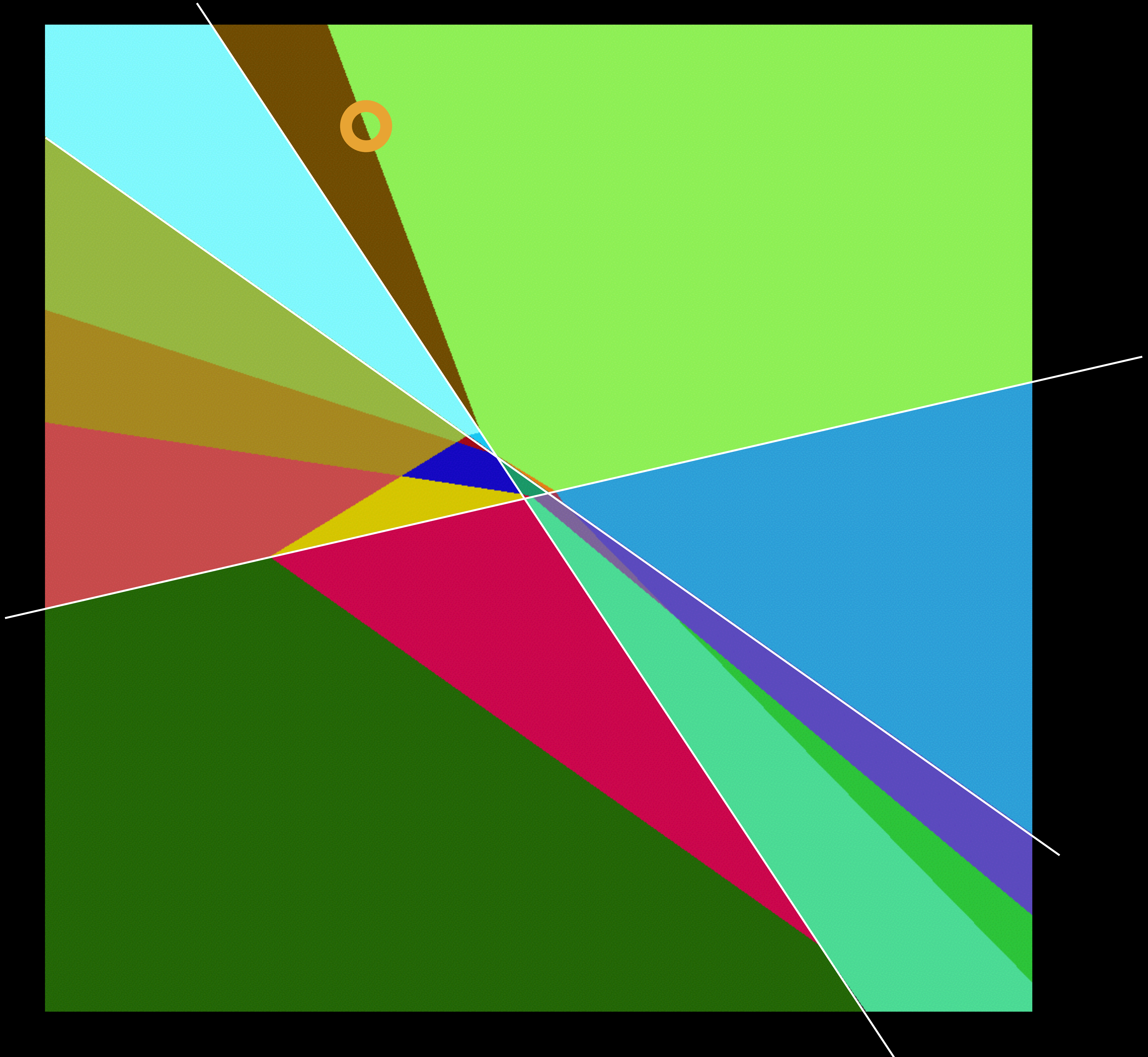


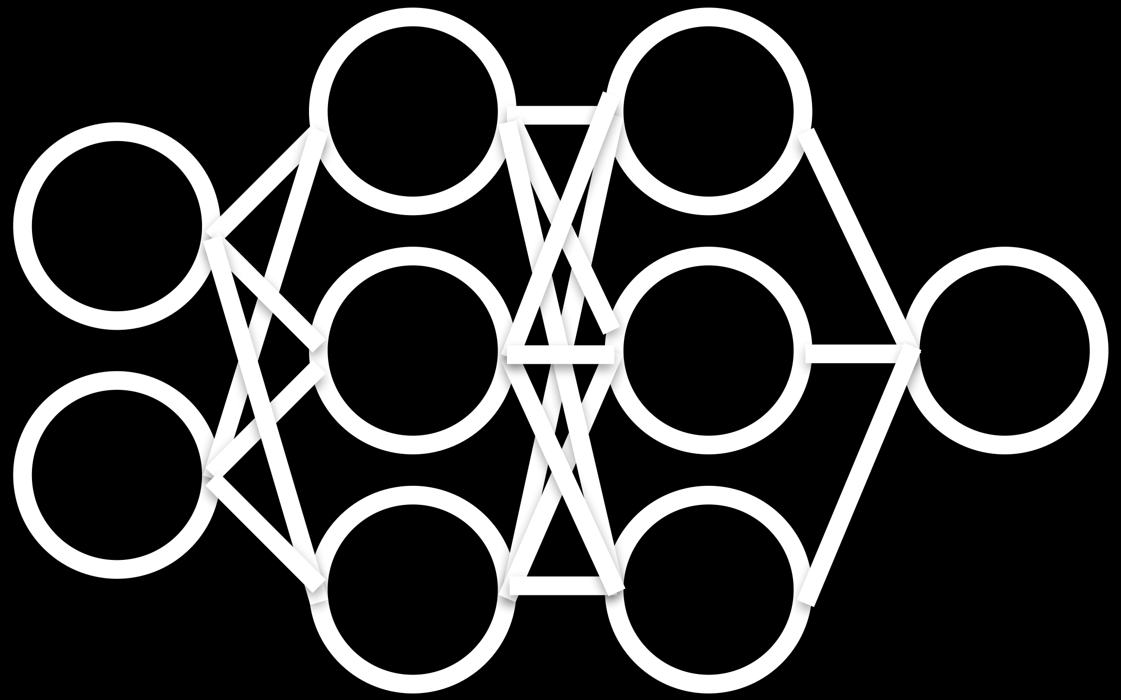


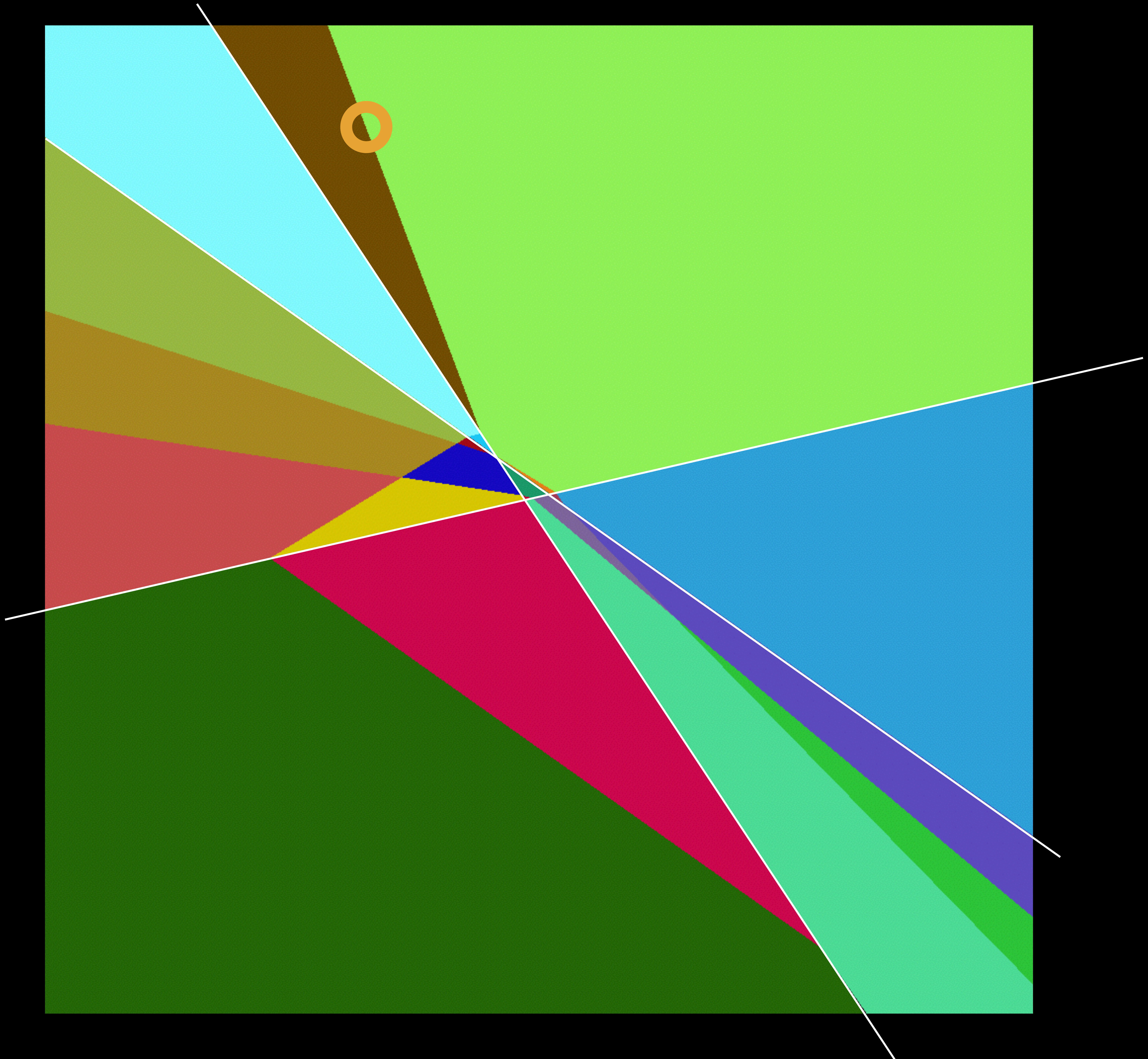
Recovering the

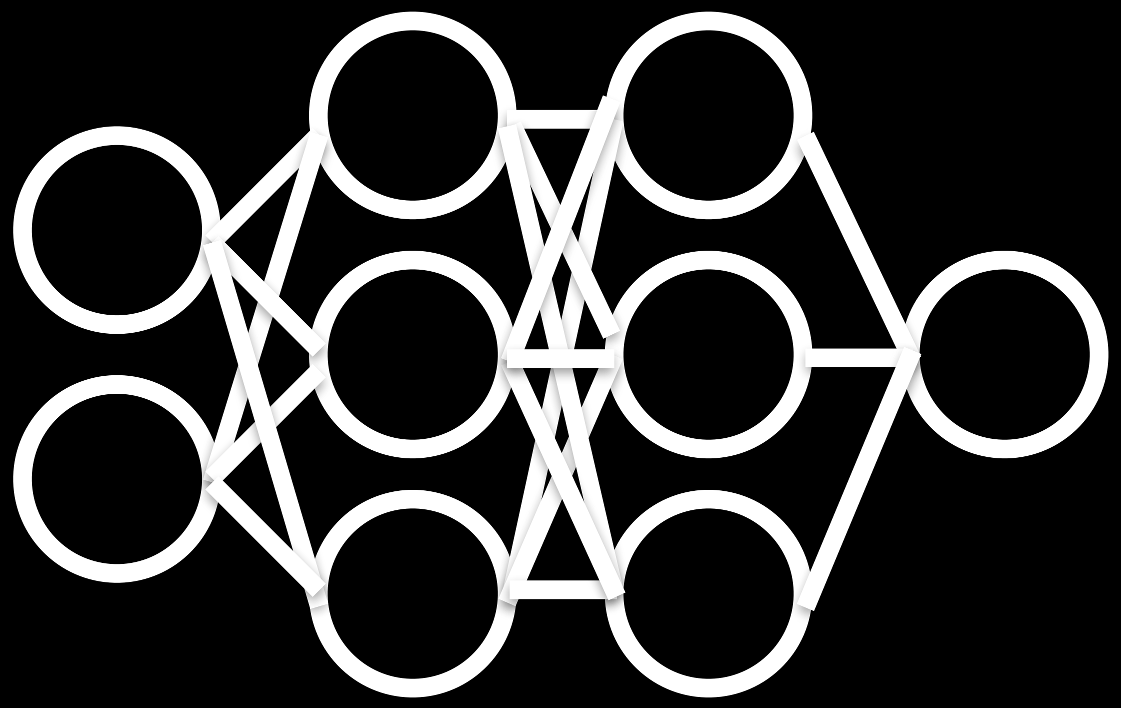
first layer sign

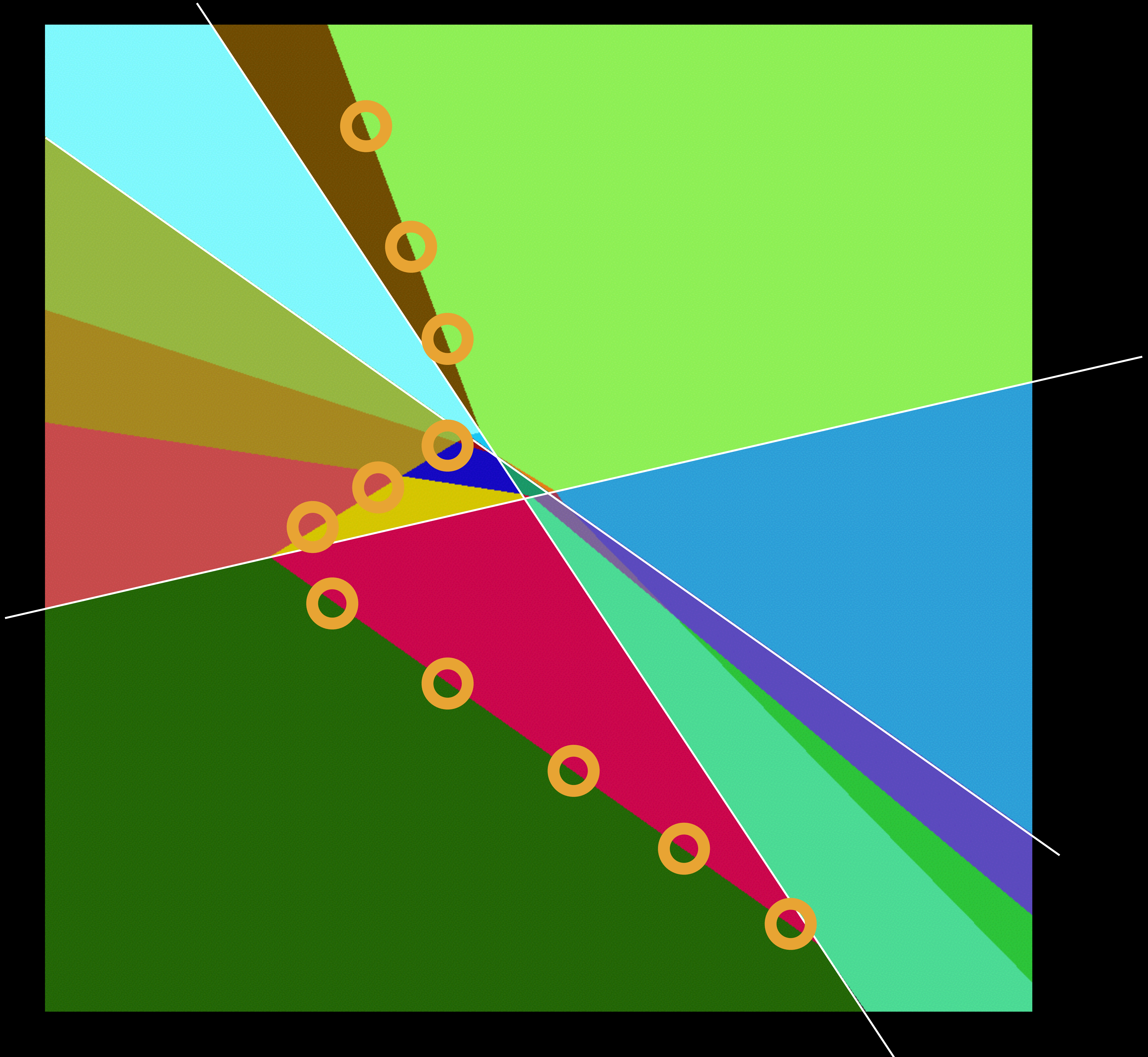


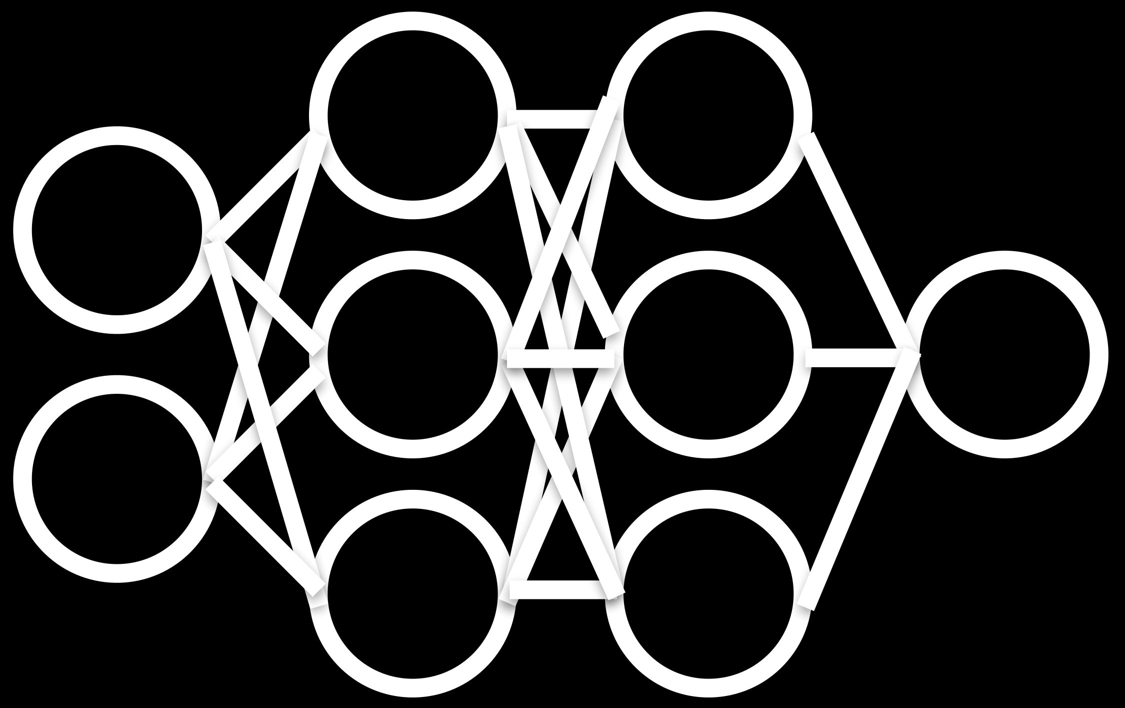


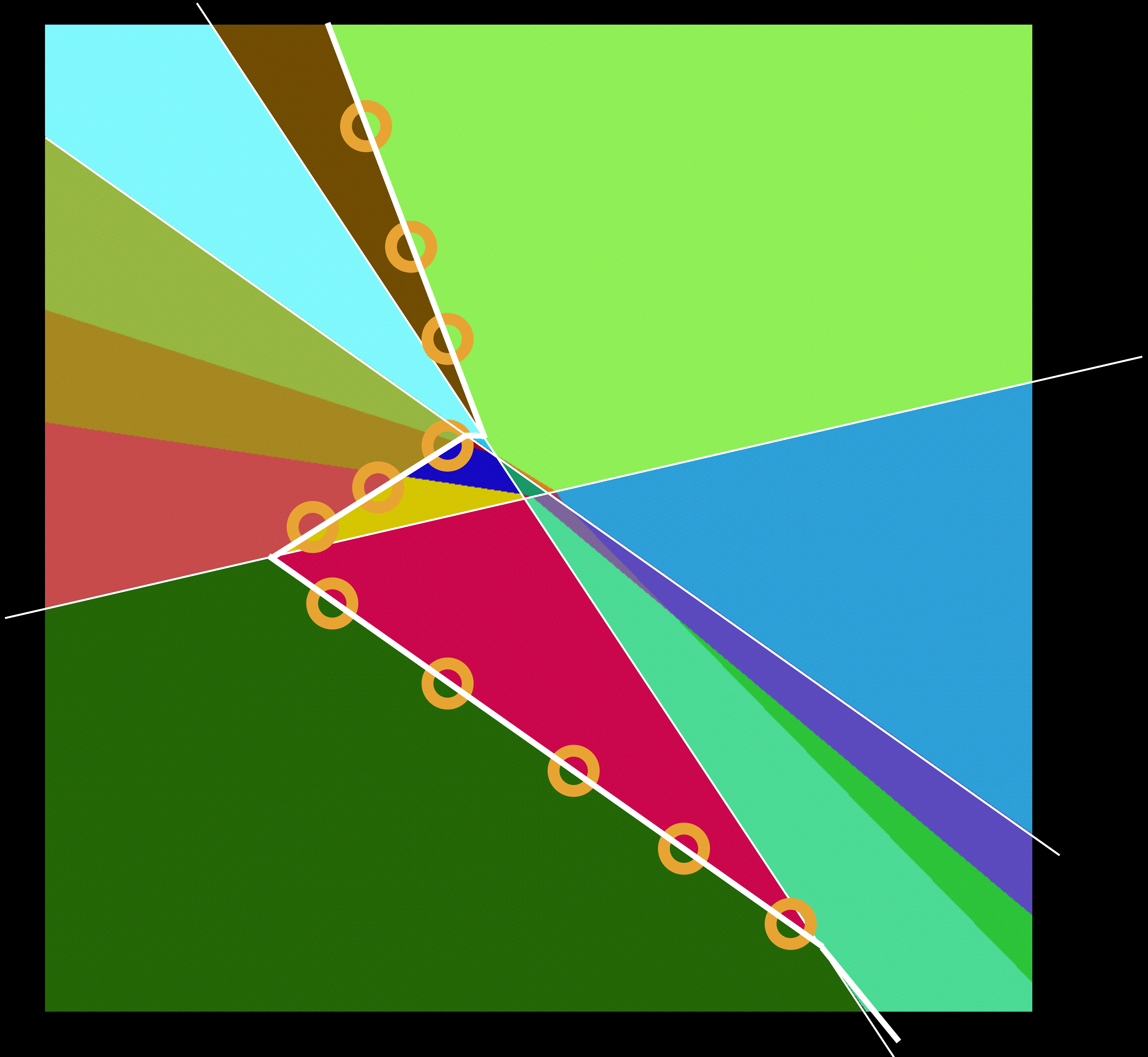


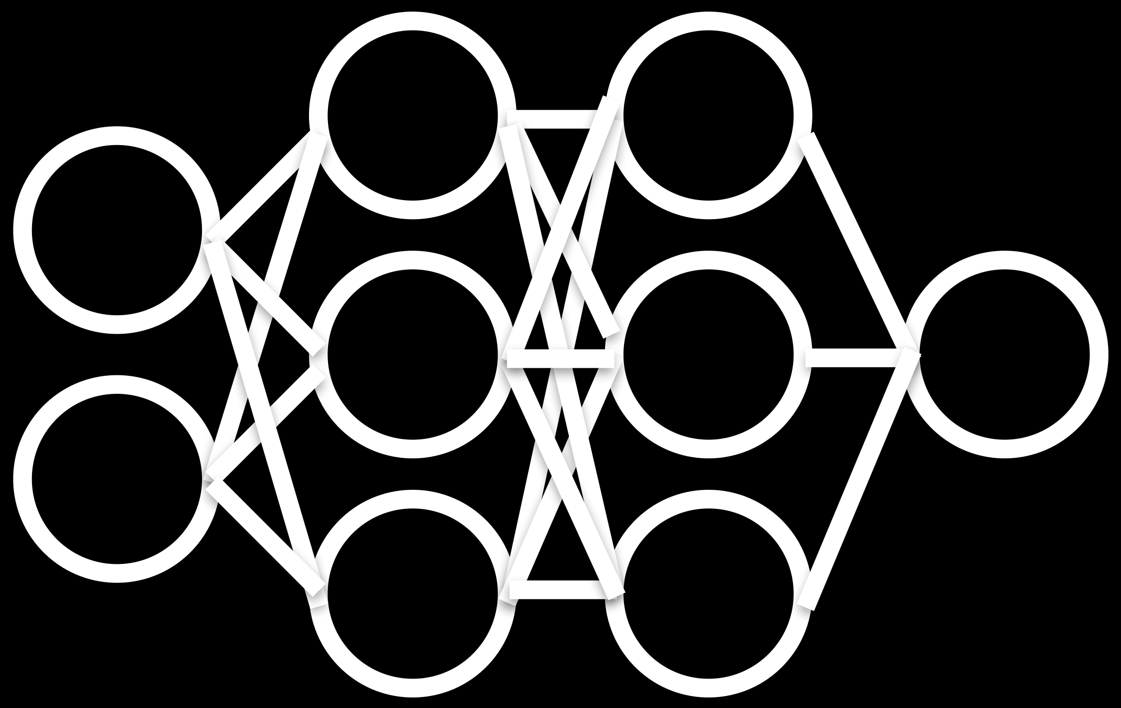


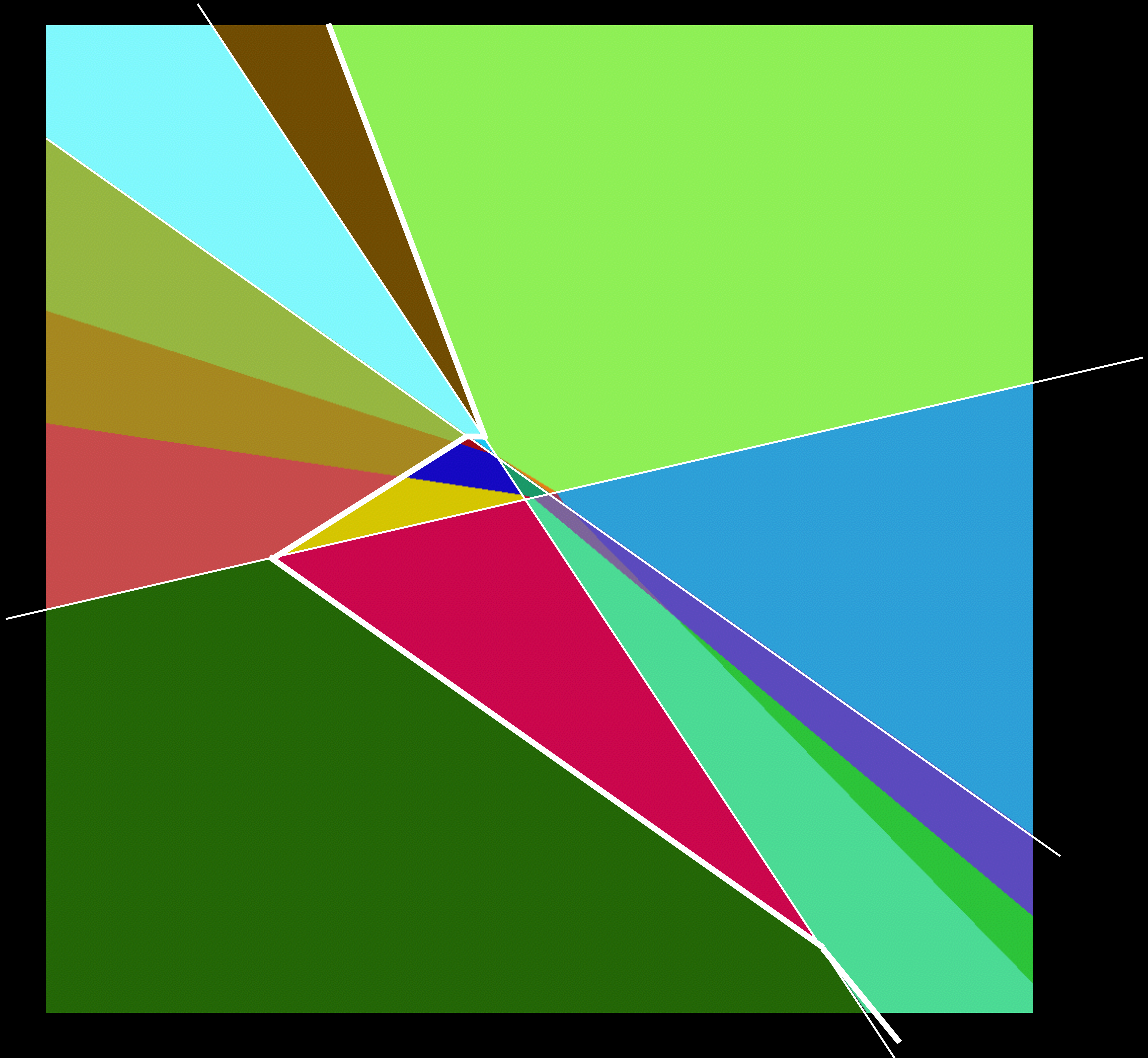


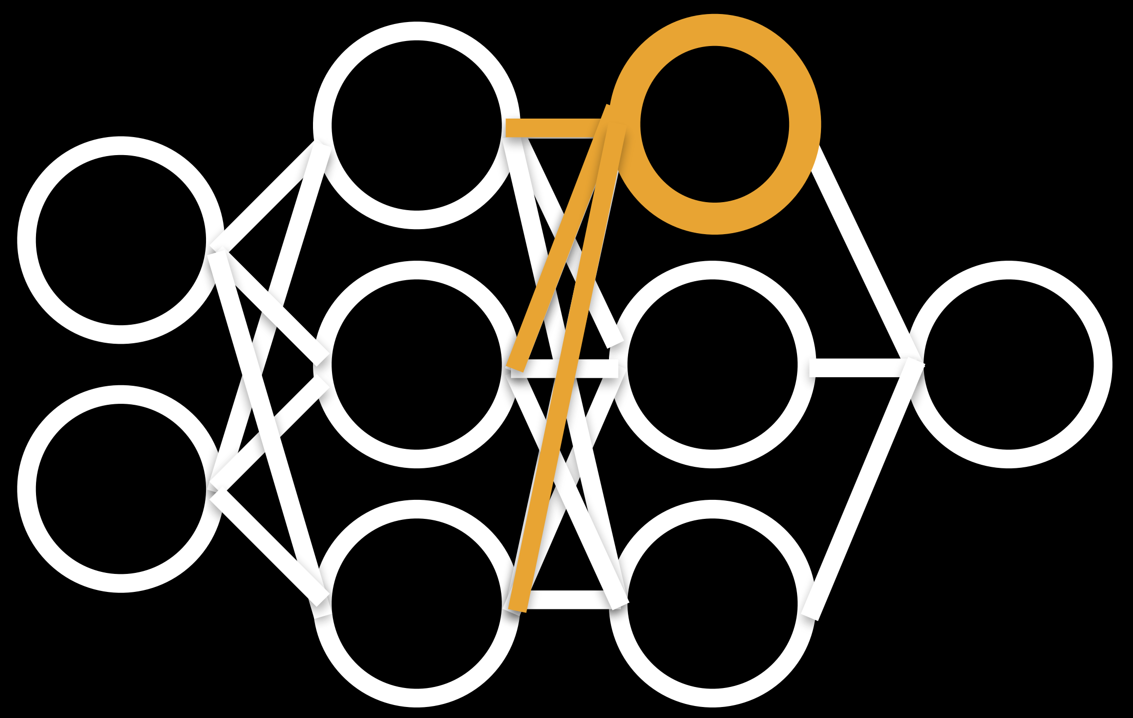


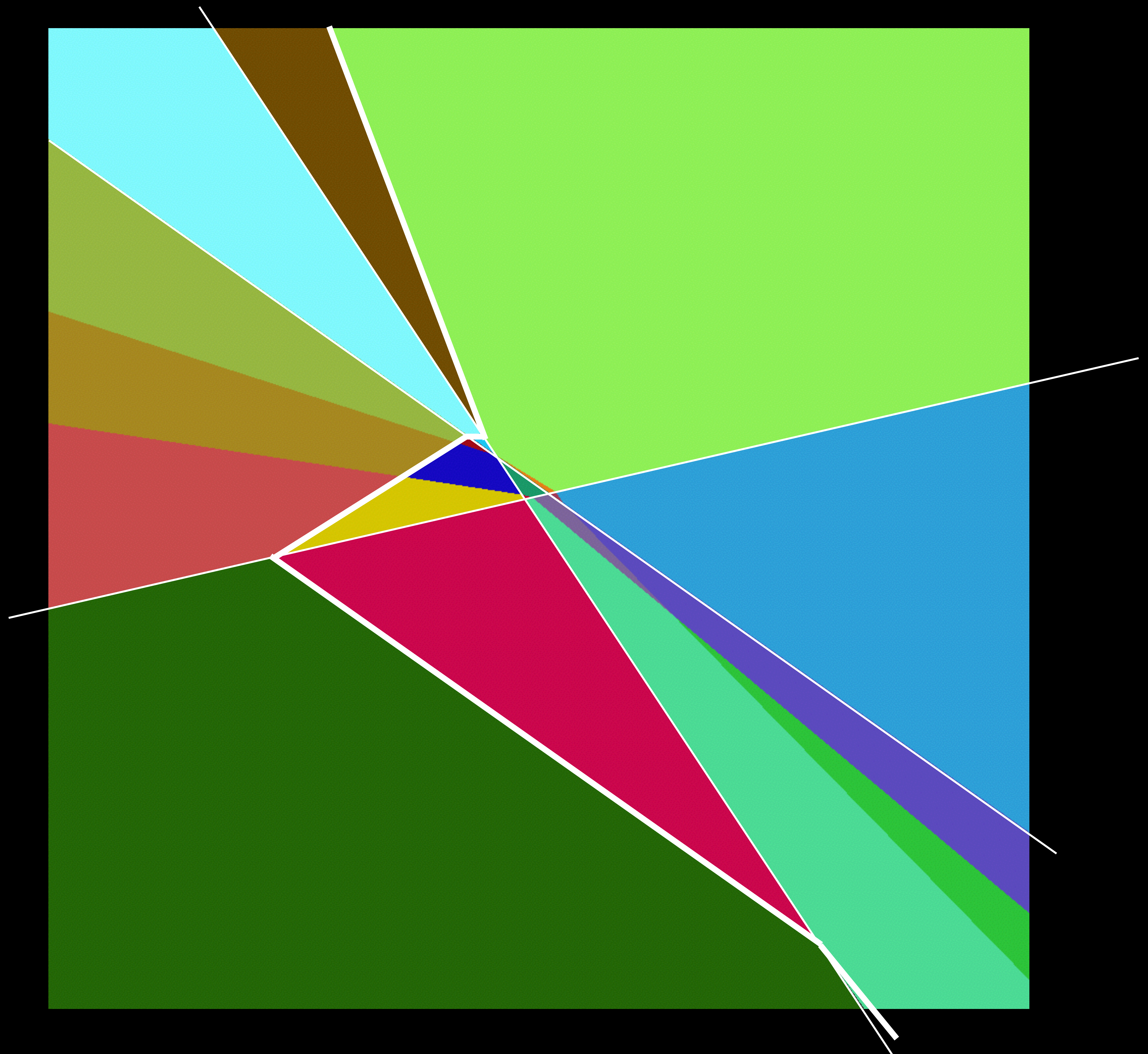




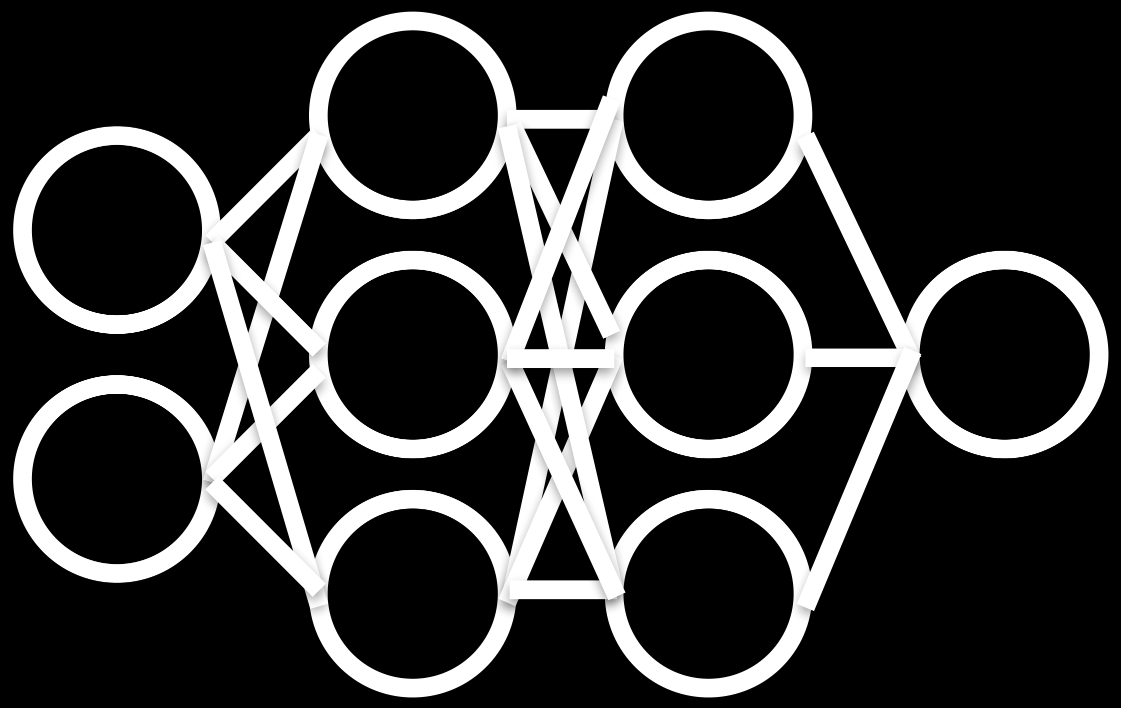


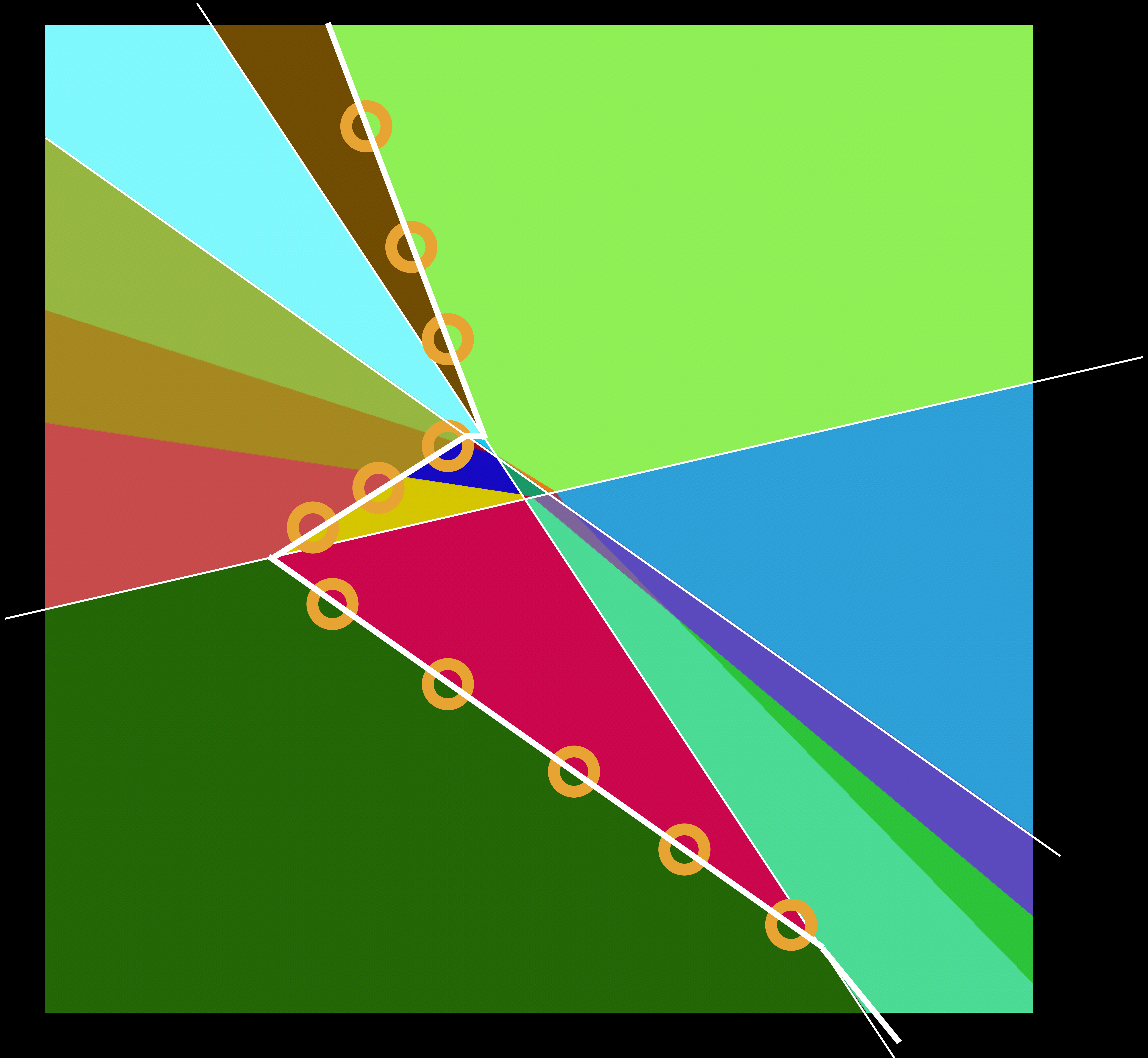


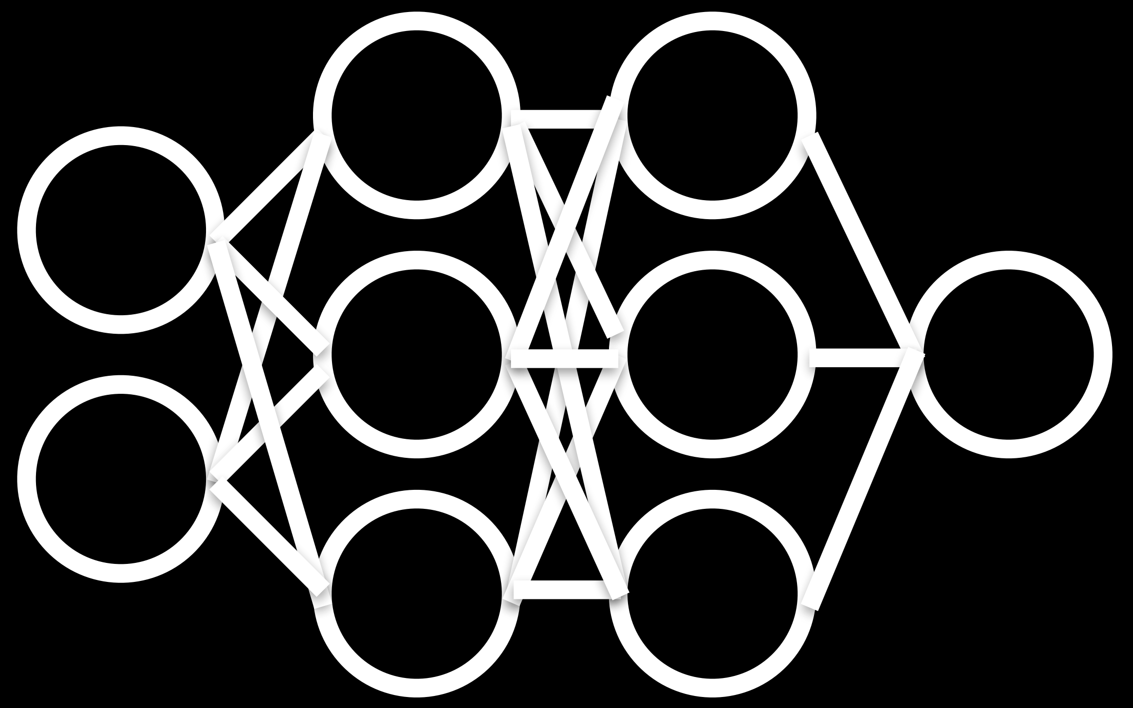


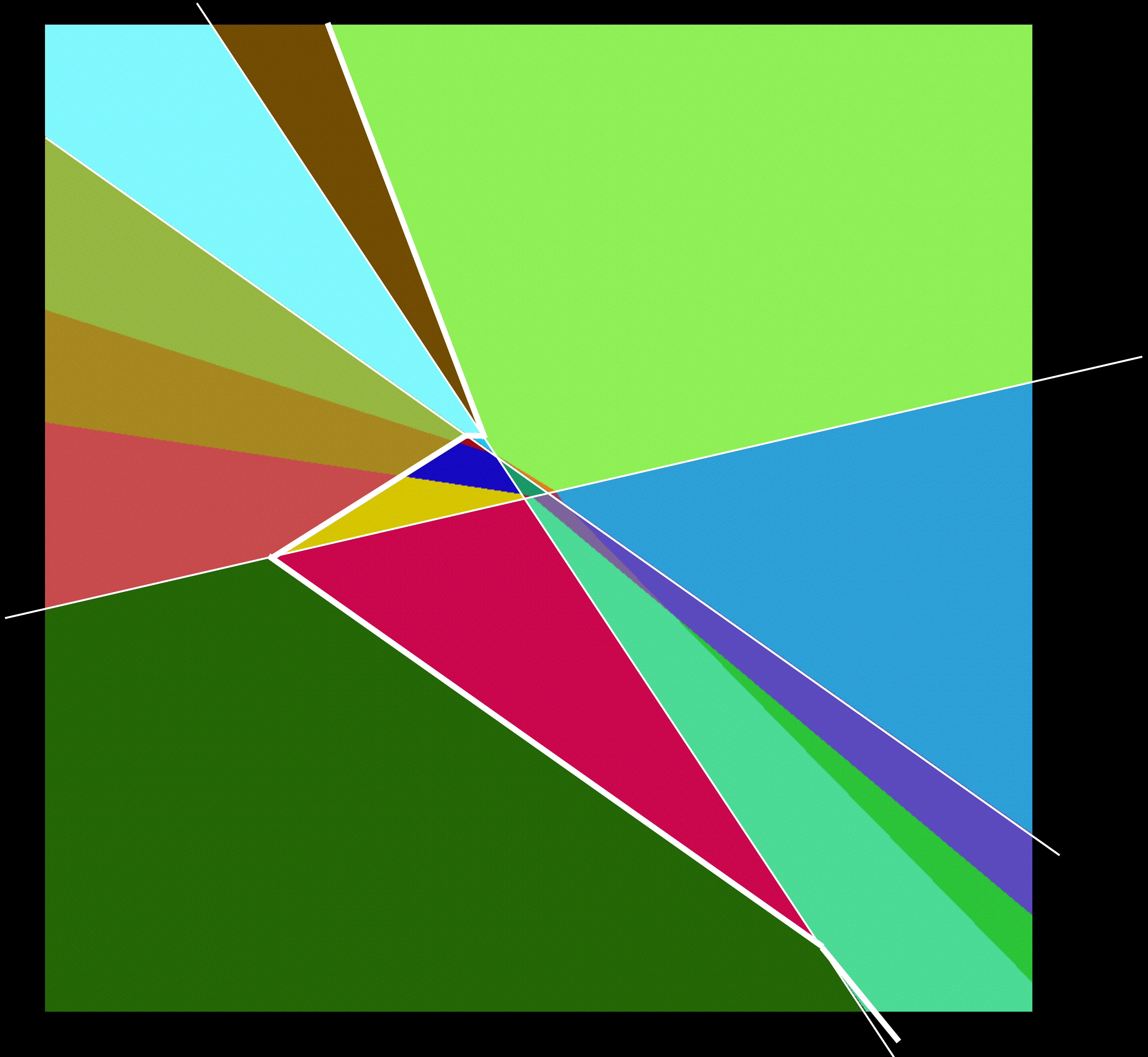


Hyperplane Following









… then peel off first weight,
and re-run attack from there

Key Challenges

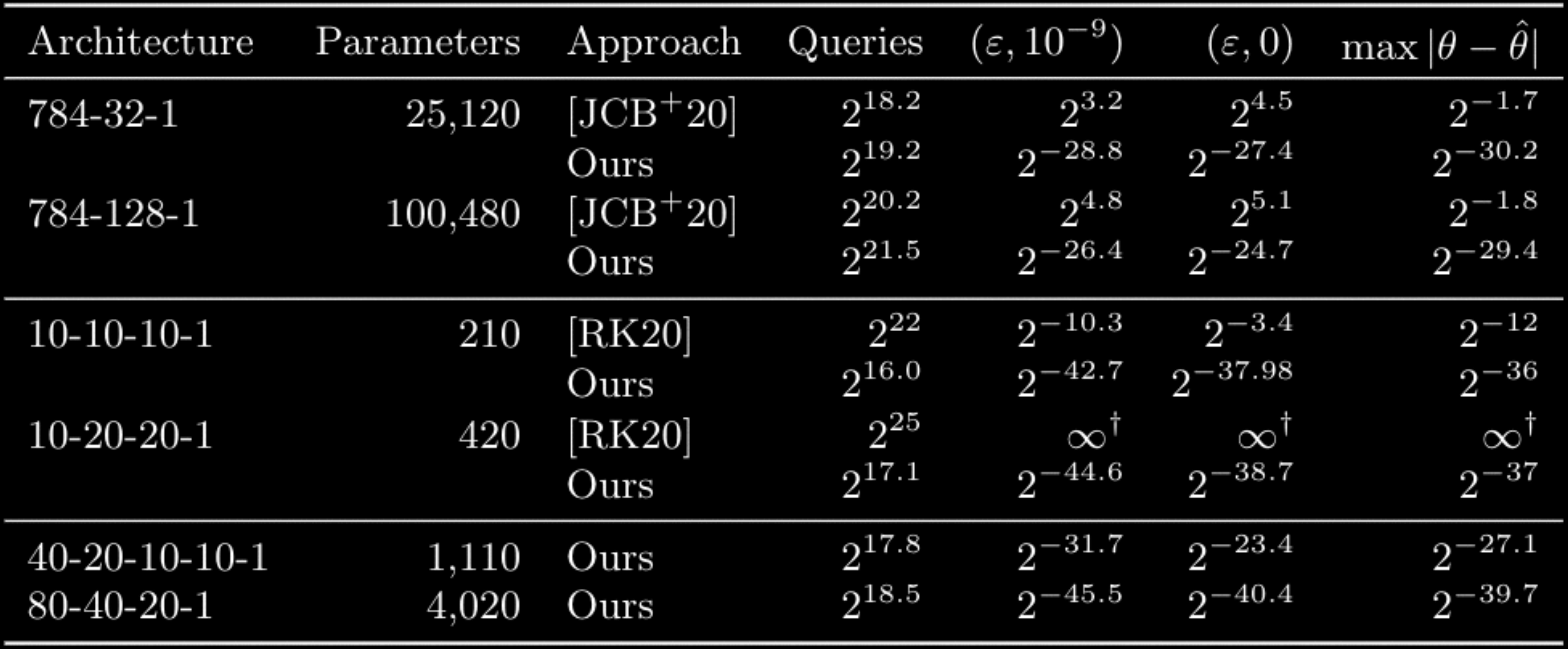
(that I don't have time to discuss in this talk, but make up most of the technical work that we had to do in the paper)

Bounded floating point precision

Not all hidden states are reachable

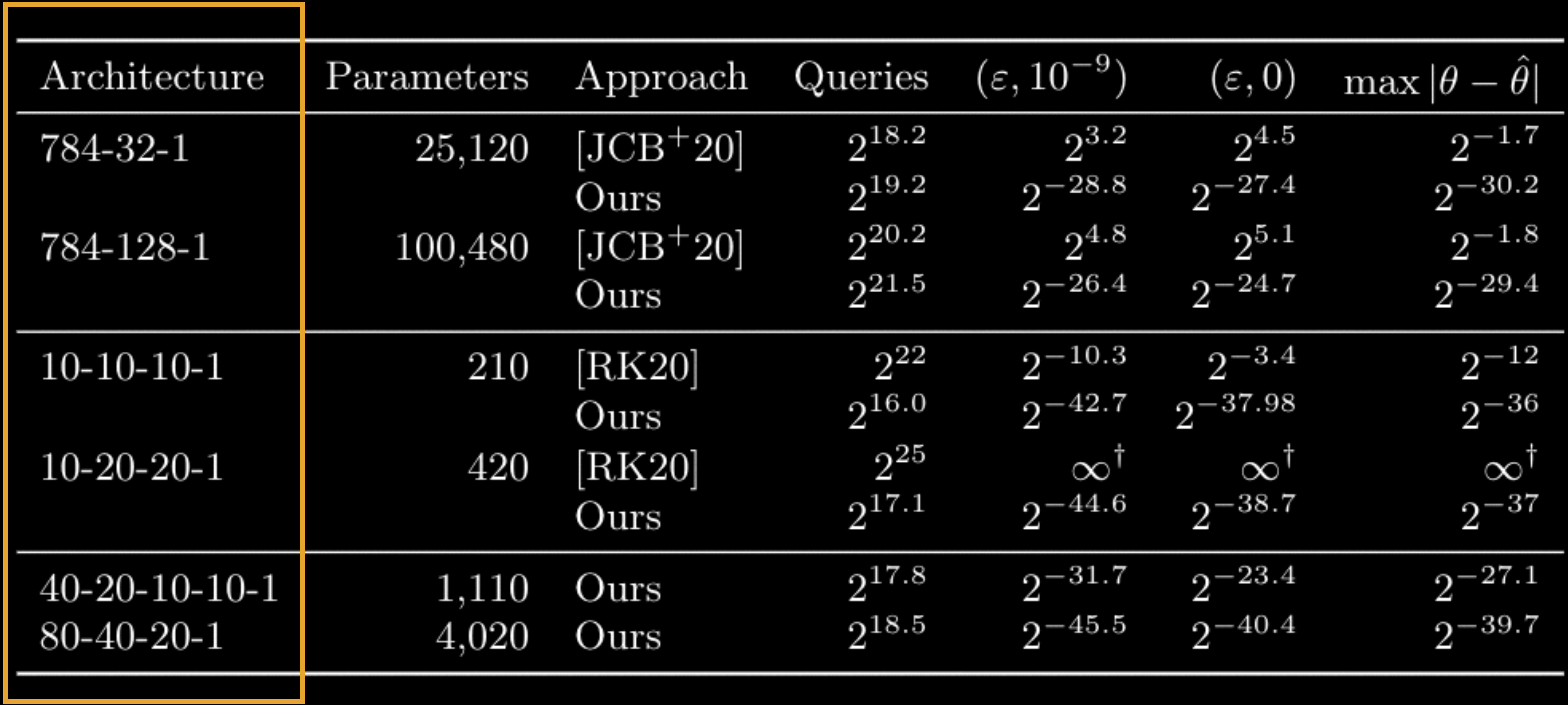
Results

Results



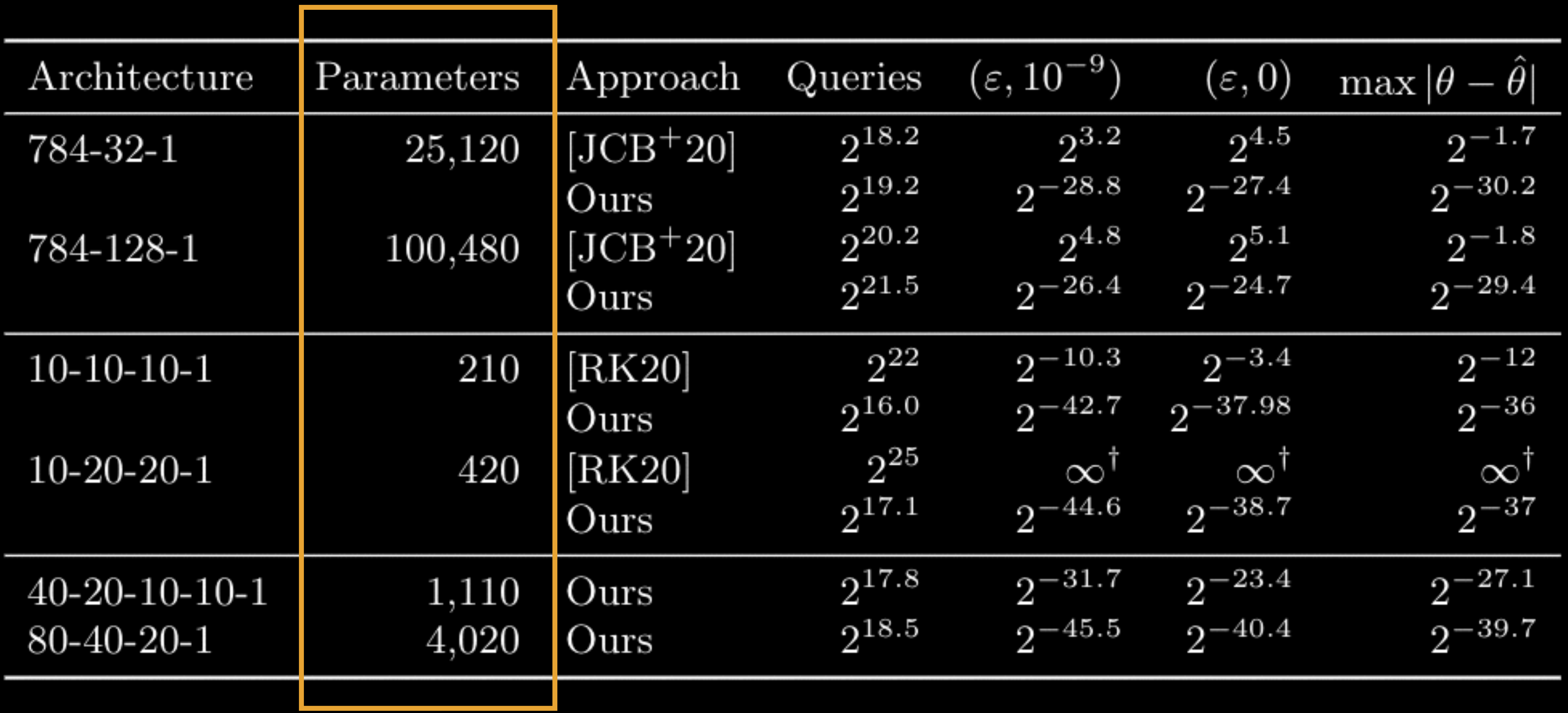
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max\|\theta - \hat{\theta}\|$
784-32-1	25,120	[JCB$^+$20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB$^+$20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^\dagger$	$\infty^\dagger$	$\infty^\dagger$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



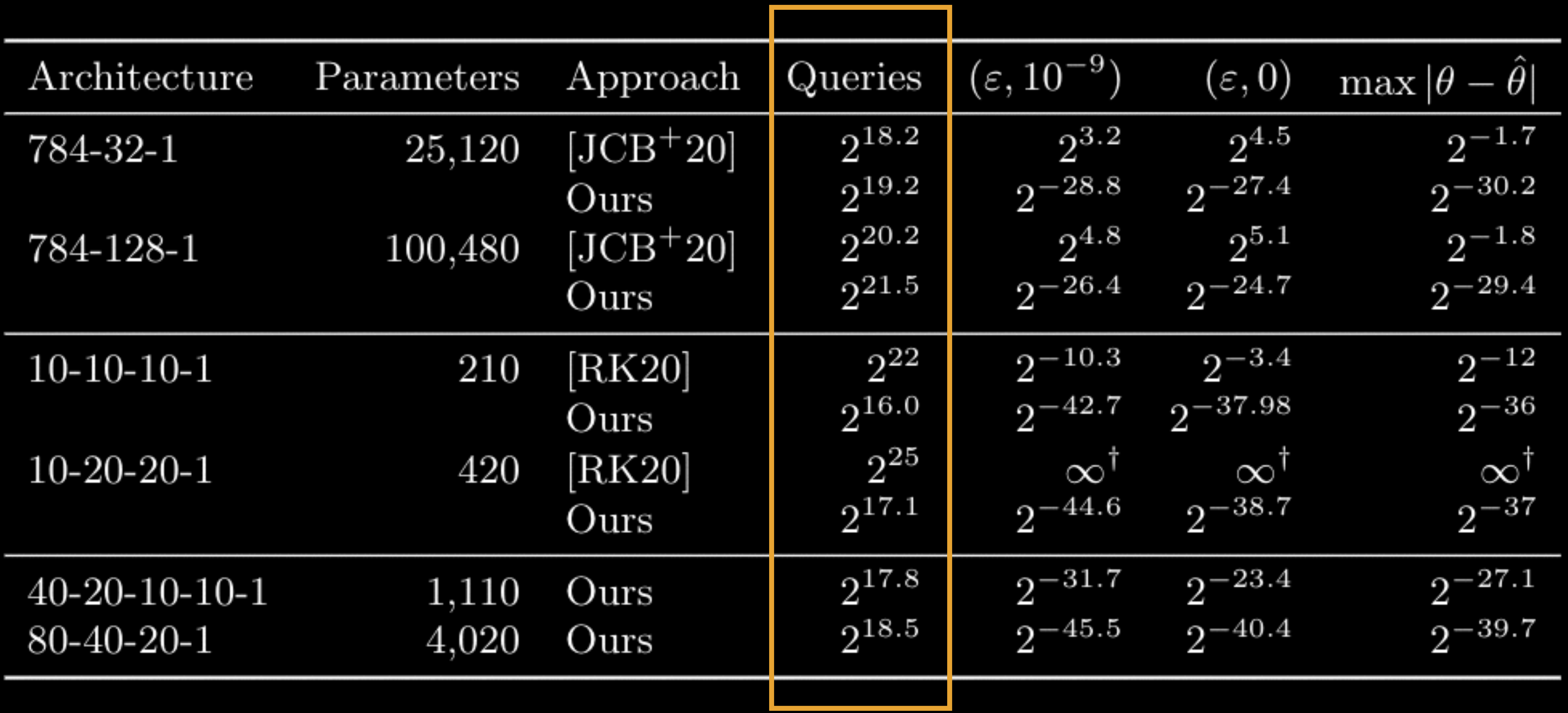
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max\lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB$^+$20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB$^+$20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^\dagger$	$\infty^\dagger$	$\infty^\dagger$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



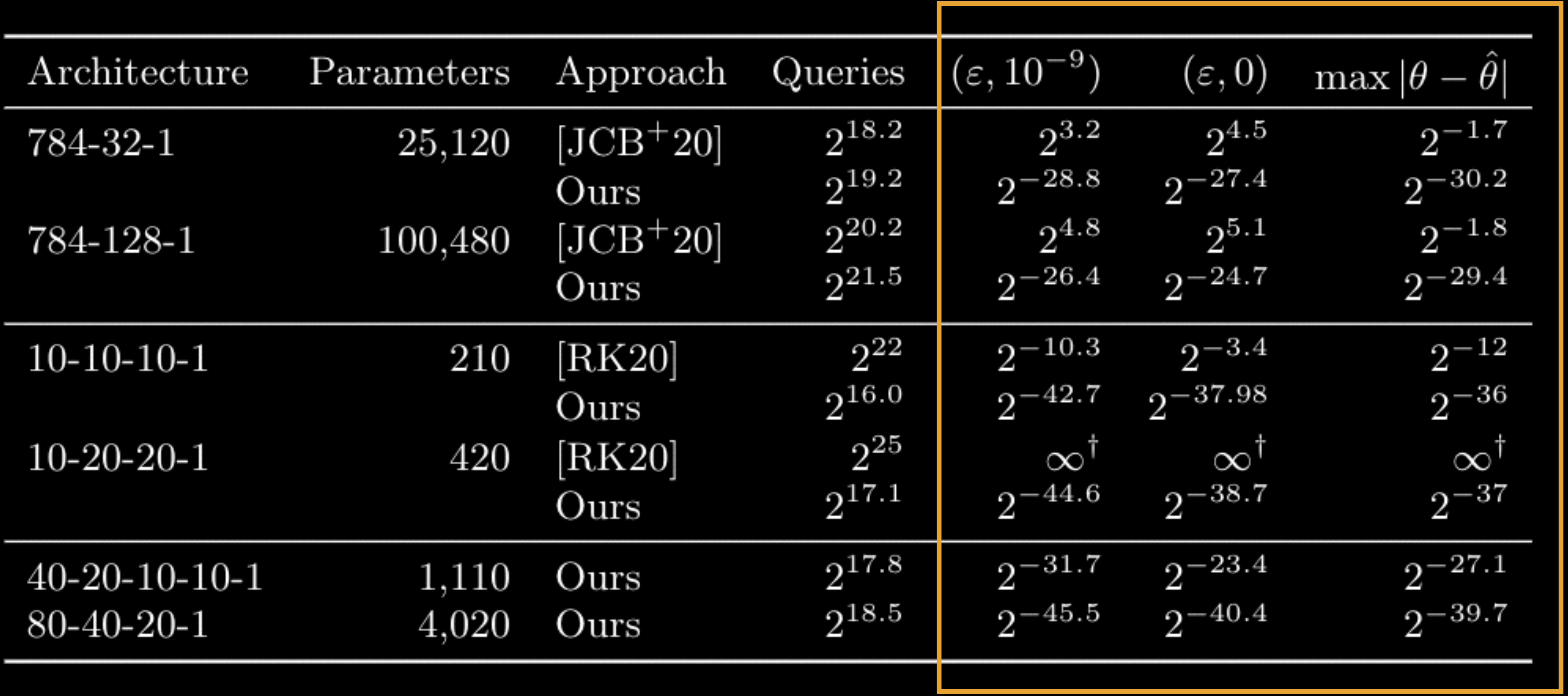
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert \theta - \hat{\theta} \rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results

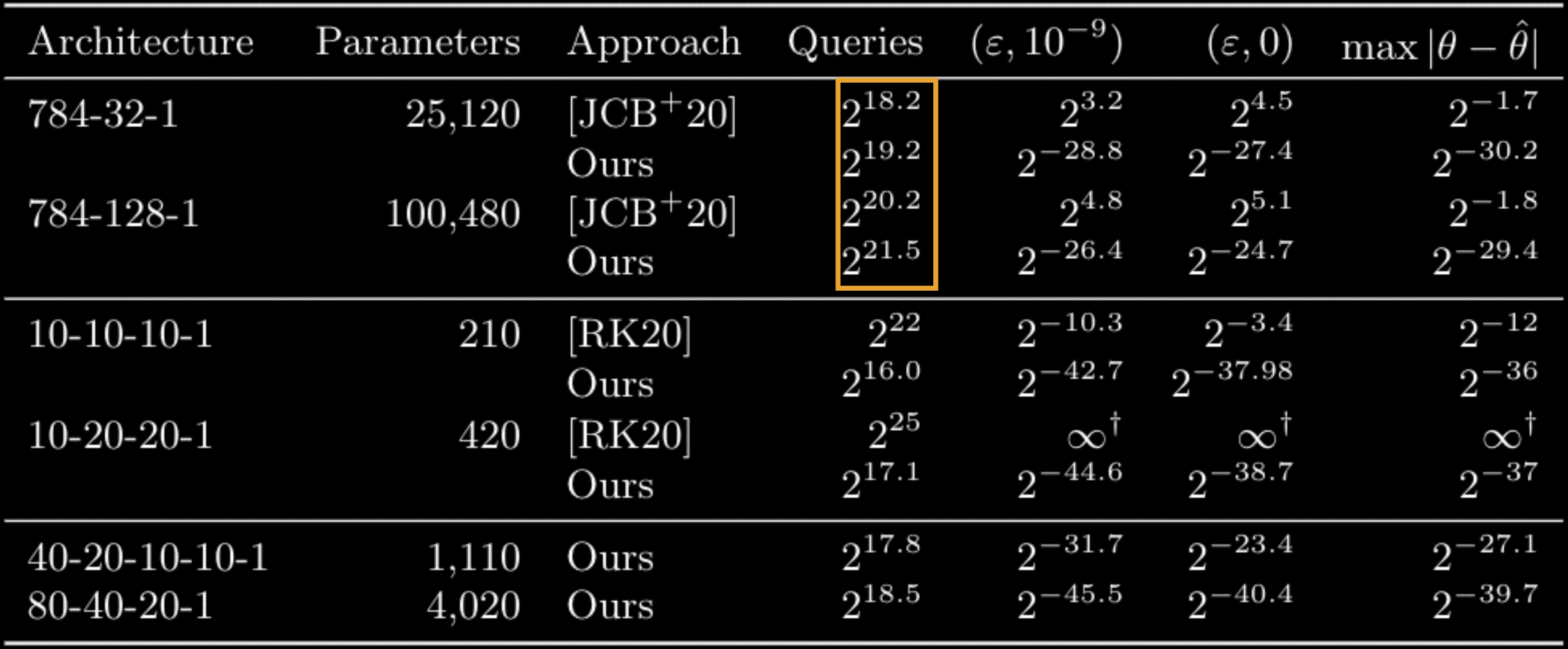


Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results

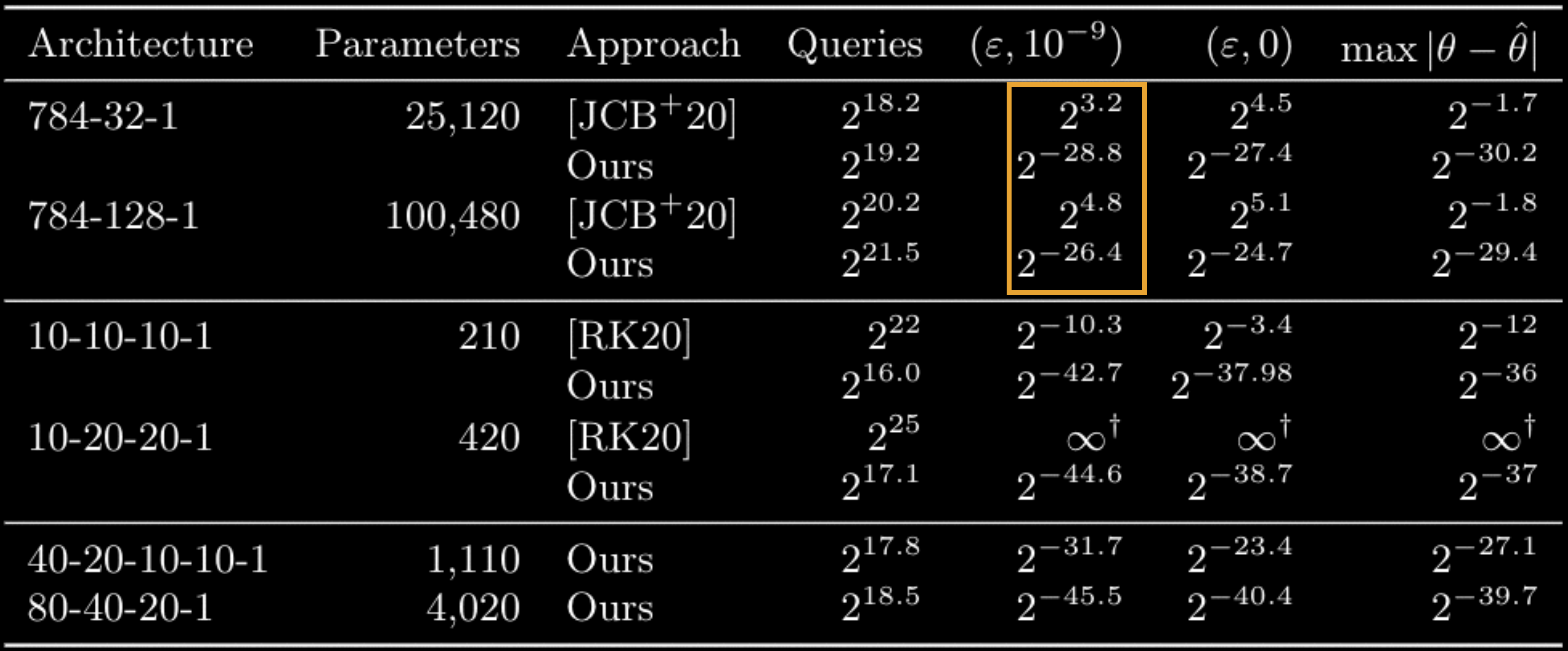
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



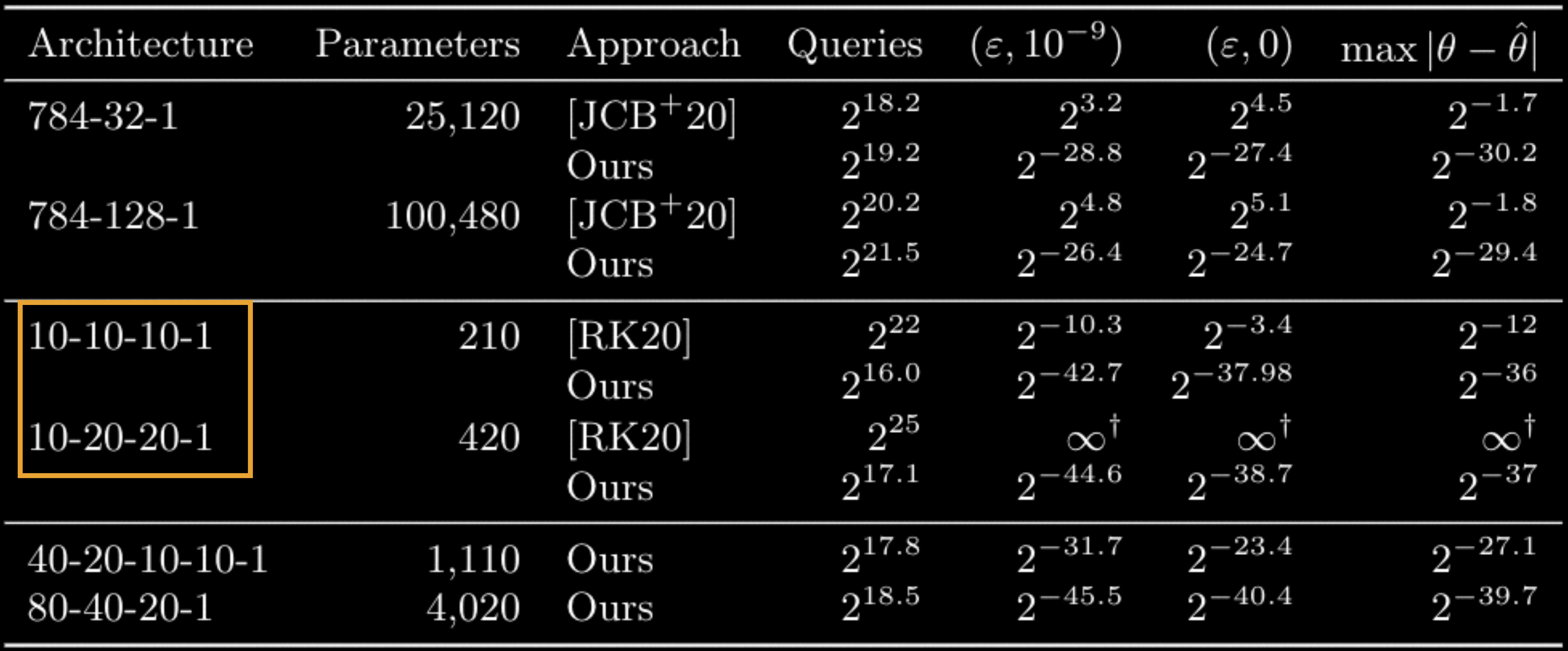
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



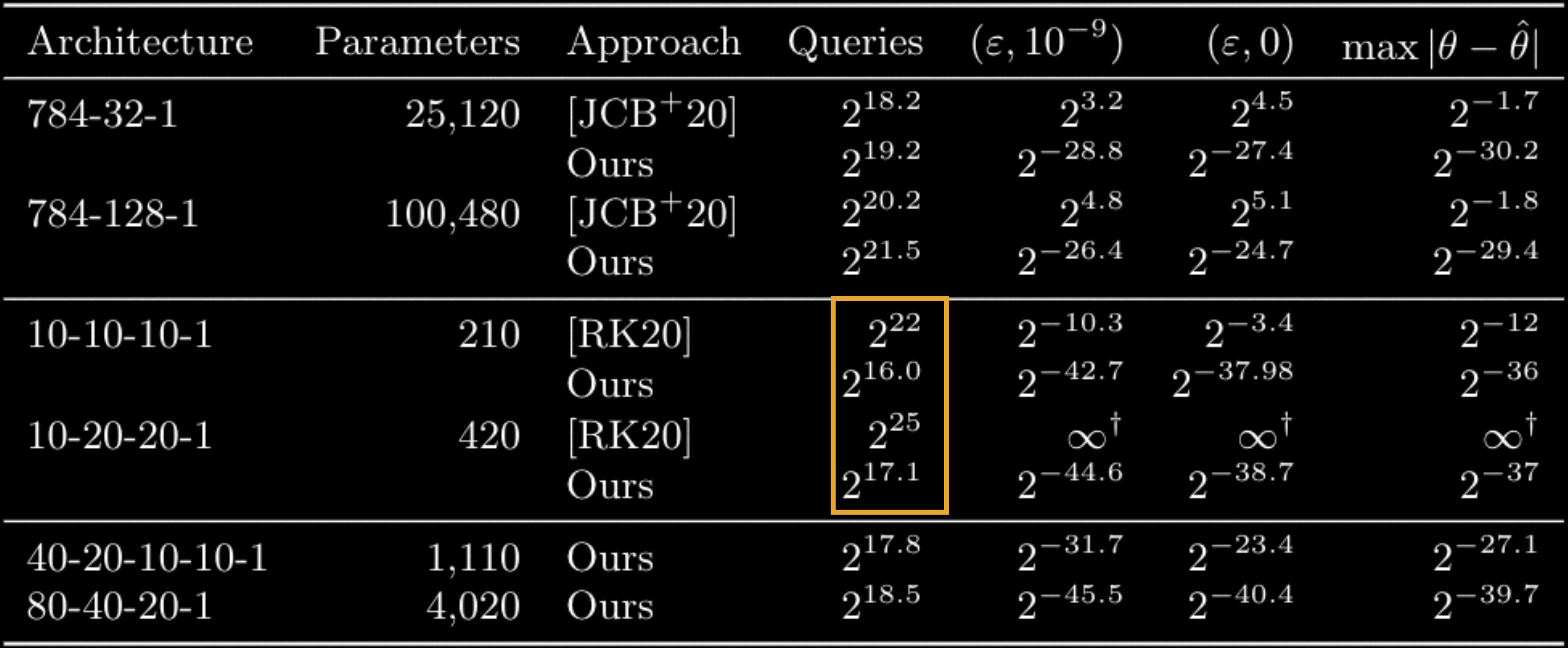
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



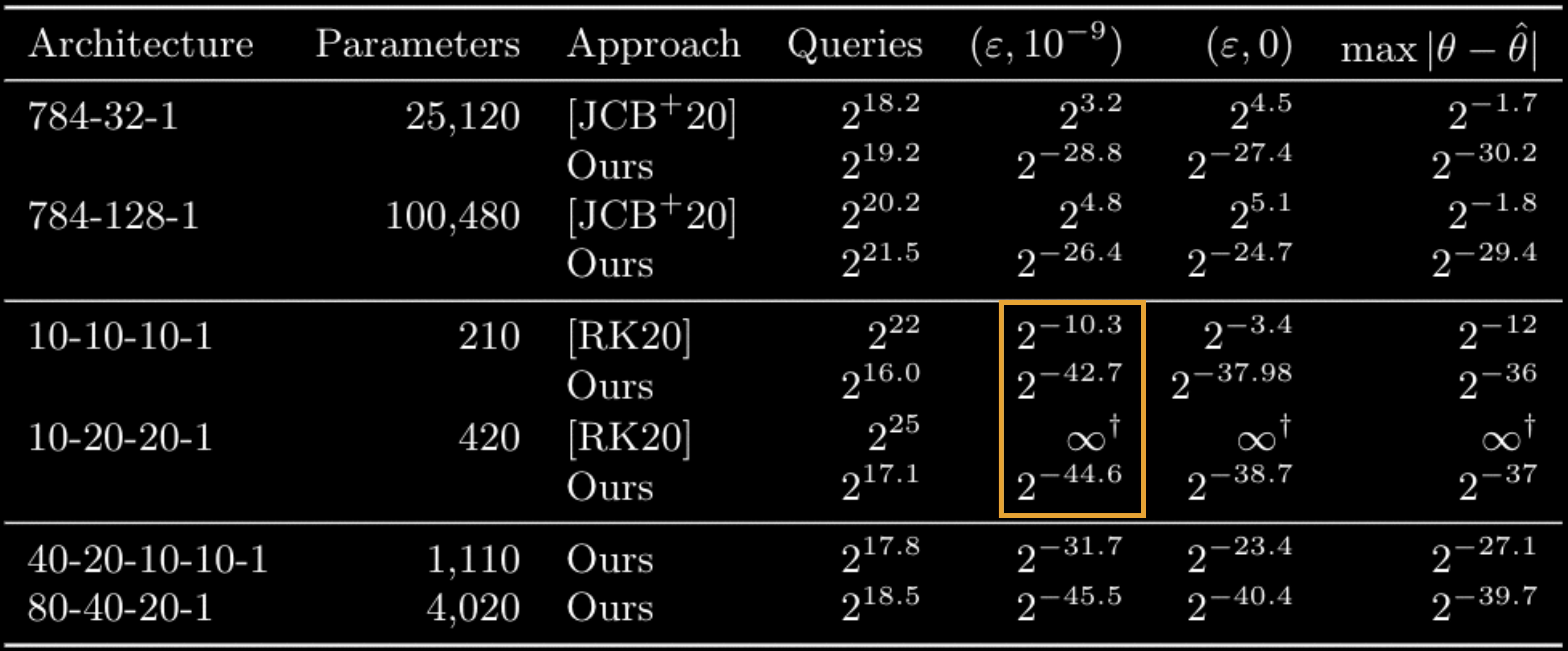
Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max\lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB$^+$20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB$^+$20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^\dagger$	$\infty^\dagger$	$\infty^\dagger$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max\lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Results



Architecture	Parameters	Approach	Queries	$(\varepsilon, 10^{-9})$	$(\varepsilon, 0)$	$\max \lvert\theta - \hat{\theta}\rvert$
784-32-1	25,120	[JCB+20]	$2^{18.2}$	$2^{3.2}$	$2^{4.5}$	$2^{-1.7}$
		Ours	$2^{19.2}$	$2^{-28.8}$	$2^{-27.4}$	$2^{-30.2}$
784-128-1	100,480	[JCB+20]	$2^{20.2}$	$2^{4.8}$	$2^{5.1}$	$2^{-1.8}$
		Ours	$2^{21.5}$	$2^{-26.4}$	$2^{-24.7}$	$2^{-29.4}$
10-10-10-1	210	[RK20]	2^{22}	$2^{-10.3}$	$2^{-3.4}$	2^{-12}
		Ours	$2^{16.0}$	$2^{-42.7}$	$2^{-37.98}$	2^{-36}
10-20-20-1	420	[RK20]	2^{25}	$\infty^{\dagger}$	$\infty^{\dagger}$	$\infty^{\dagger}$
		Ours	$2^{17.1}$	$2^{-44.6}$	$2^{-38.7}$	2^{-37}
40-20-10-10-1	1,110	Ours	$2^{17.8}$	$2^{-31.7}$	$2^{-23.4}$	$2^{-27.1}$
80-40-20-1	4,020	Ours	$2^{18.5}$	$2^{-45.5}$	$2^{-40.4}$	$2^{-39.7}$

Conclusions

Direct analysis of neural networks

"Secure Inference"
maybe isn't so secure ...

Don't put neural networks in your ideal functionalities

-A talk by Matthew Jagielski

Live Q&A:
Friday 8:00 PT / 15:00 UTC

After-the-fact Q&A: nicholas@carlini.com

Code: https://github.com/google-research/cryptanalytic-model-extraction

Live Q&A:

Friday 8:00 PT / 15:00 UTC

After-the-fact Q&A: nicholas@carlini.com

Code: https://github.com/google-research/cryptanalytic-model-extraction