Generalization in Neural Networks
Through the Lens of Magnitude Potential

Sahel Torkamani · Henry Gouk · Rik Sarkar

Informatics, University of Edinburgh



THE UNIVERSITY of EDINBURGH

NeurIPS 2026

December 2026

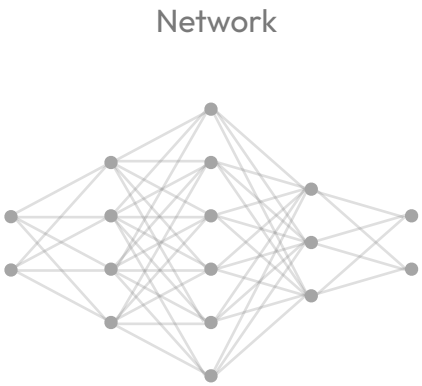
ArXiv

Homepage

Motivation

How Networks Treat Their Data

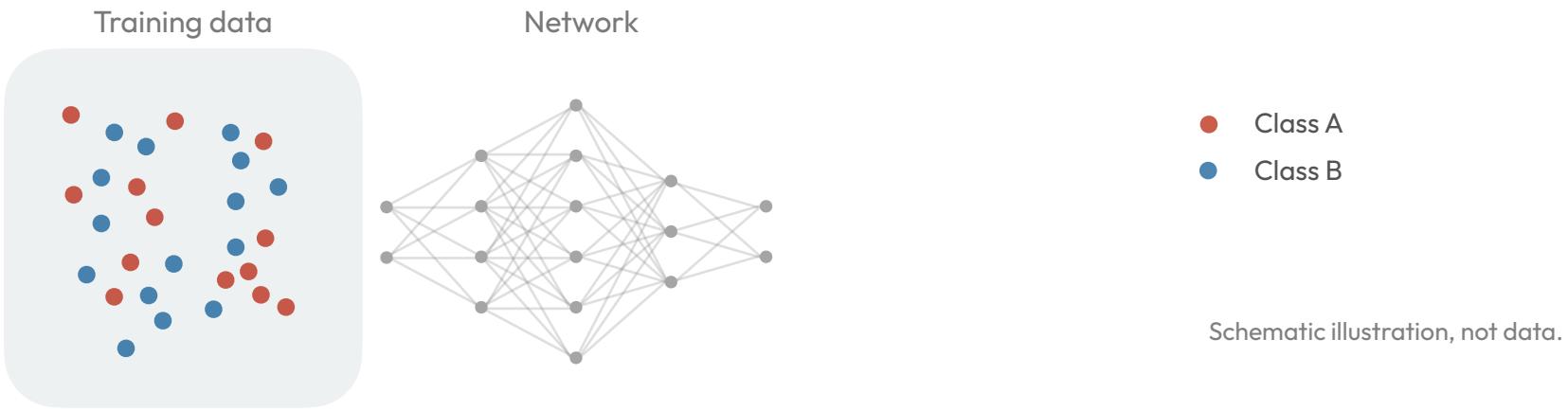
Looking at what a **network** does with each training sample in their logit layer, we want to answer:



Motivation

How Networks Treat Their Data

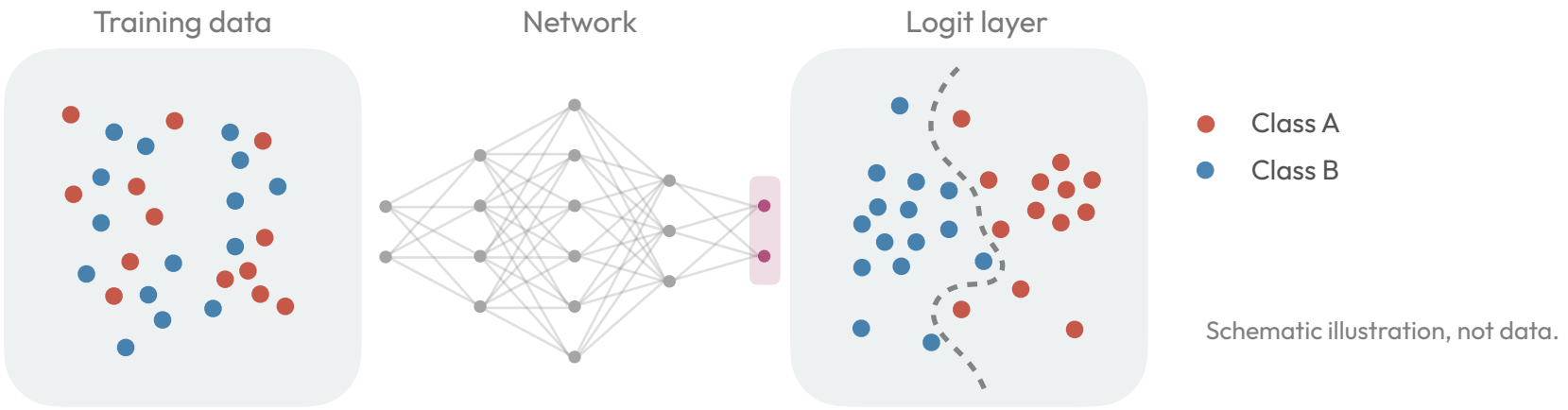
Looking at what a network does with each **training sample** in their logit layer, we want to answer:



Motivation

How Networks Treat Their Data

Looking at what a network does with each training sample in **their logit layer,** we want to answer:

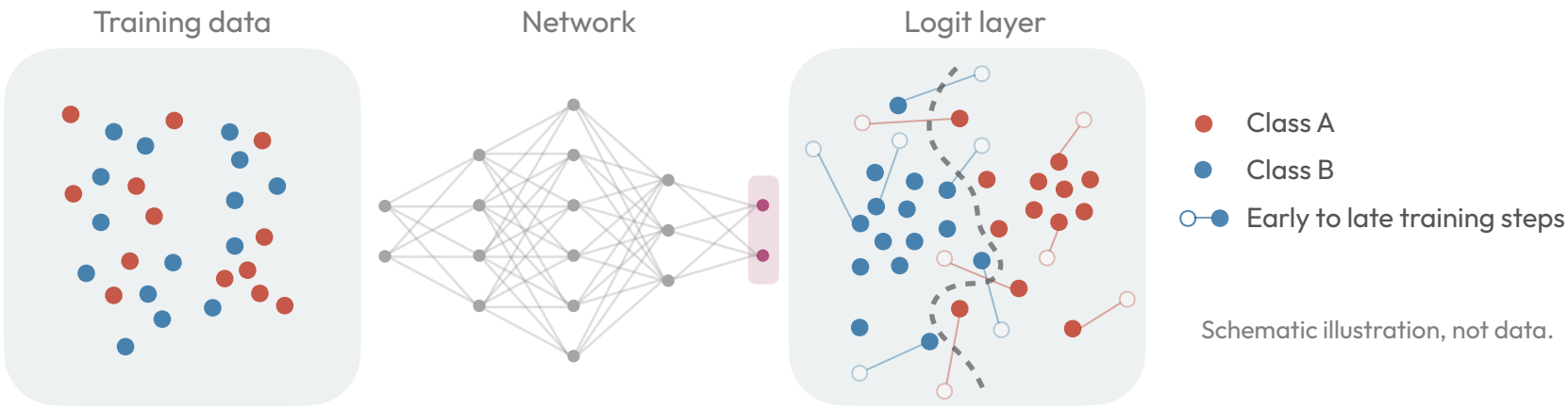


Motivation

How Networks Treat Their Data

Looking at what a network does with each training sample in their logit layer, we want to answer:

- How does the representation of points change during training?

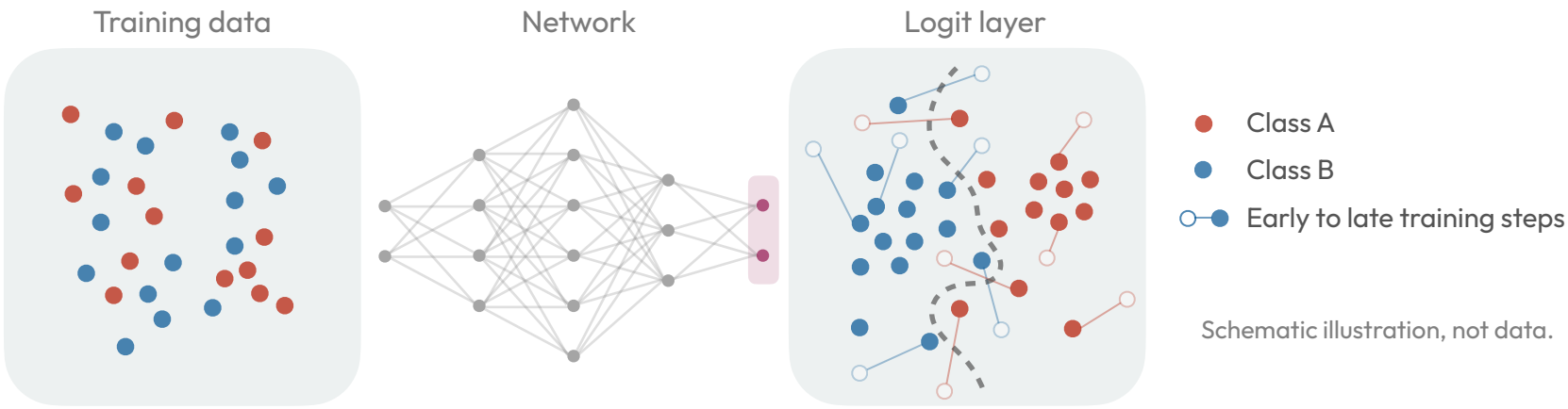


Motivation

How Networks Treat Their Data

Looking at what a network does with each training sample in their logit layer, we want to answer:

- How does the representation of points change during training?
  - How can we explain the training dynamics?

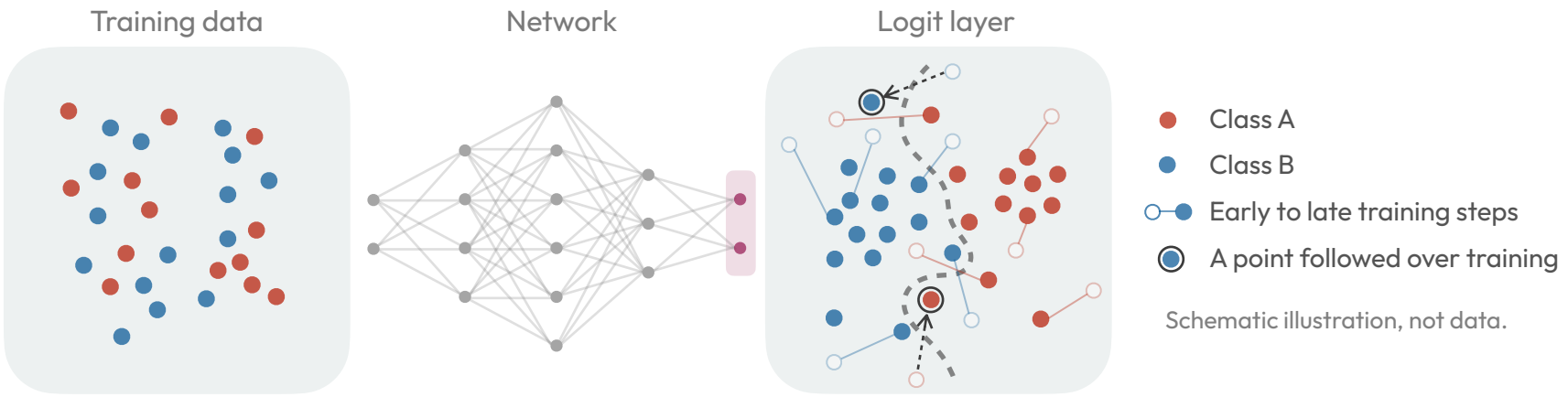


Motivation

How Networks Treat Their Data

Looking at what a network does with each training sample in their logit layer, we want to answer:

- How does the representation of points change during training?
  - How can we explain the training dynamics?
- What does a network do with each point of its training data?

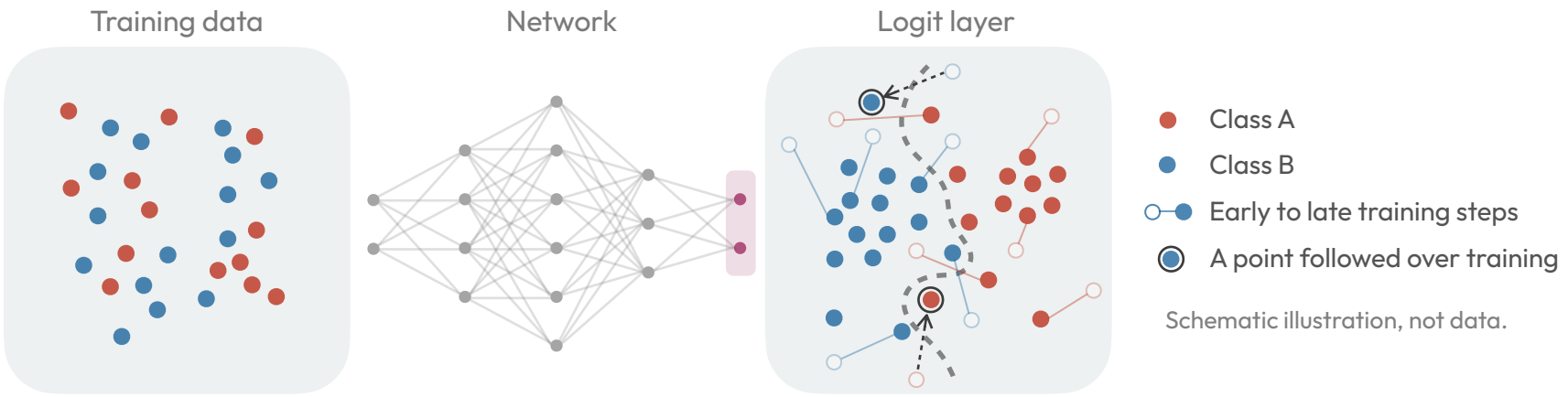


Motivation

How Networks Treat Their Data

Looking at what a network does with each training sample in their logit layer, we want to answer:

- How does the representation of points change during training?
  - How can we explain the training dynamics?
- What does a network do with each point of its training data?
  - How can we predict memorization and generalization?

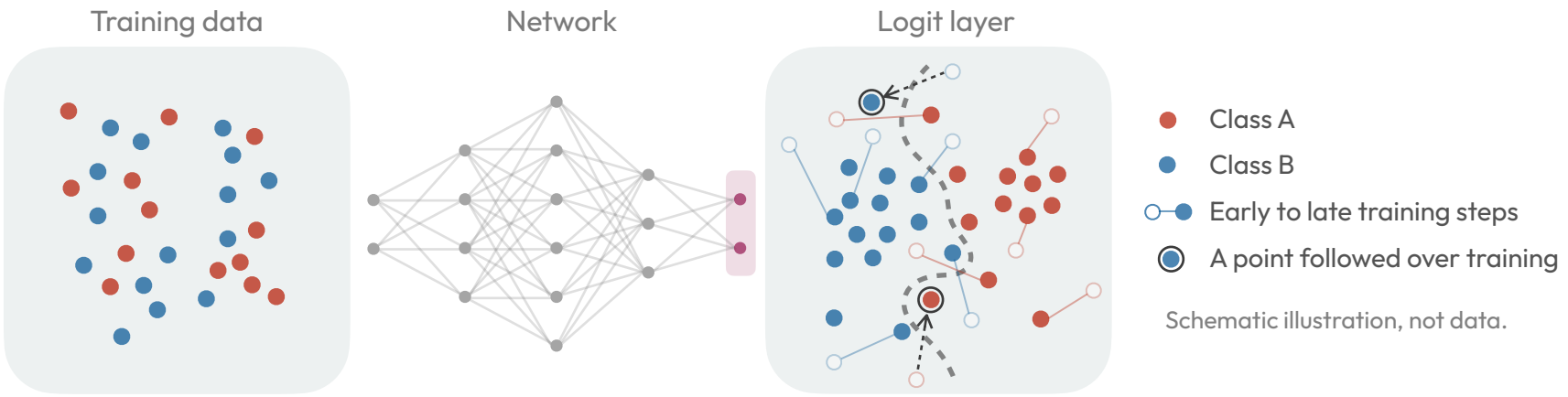


Motivation

How Networks Treat Their Data

Looking at what a network does with each training sample in their logit layer, we want to answer:

- How does the representation of points change during training?
  - How can we explain the training dynamics?
- What does a network do with each point of its training data?
  - How can we predict memorization and generalization?

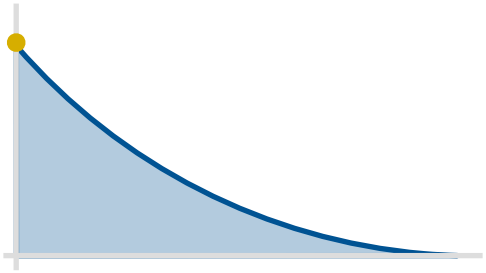


To inspect this representation, we introduce the **magnitude potential ratio**.

Background

Magnitude Potential

SIMILARITY



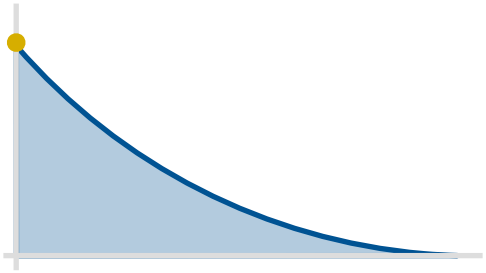
$$\zeta^t(x, y) = exp(-t \cdot d(x, y))$$

The scale t sets how quickly similarity decays with distance d.

Background

Magnitude Potential

SIMILARITY



$$\zeta^t(x, y) = exp(-t \cdot d(x, y))$$

The scale t sets how quickly similarity decays with distance d.

WEIGHTING



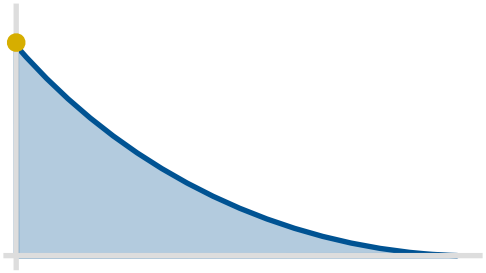
$$\mathbf{w}_X^t = (\zeta^t)^{-1}\mathbf{1}$$

The weighted similarities sum to 1 at every point of X.

Background

Magnitude Potential

SIMILARITY



$$\zeta^t(x, y) = exp(-t \cdot d(x, y))$$

The scale t sets how quickly similarity decays with distance d.

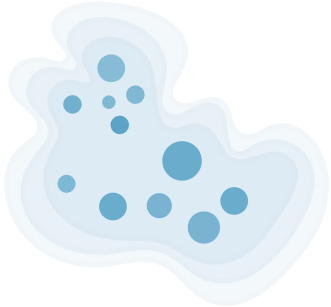
WEIGHTING



$$\mathbf{w}_X^t = (\zeta^t)^{-1}\mathbf{1}$$

The weighted similarities sum to 1 at every point of X.

MAGNITUDE POTENTIAL



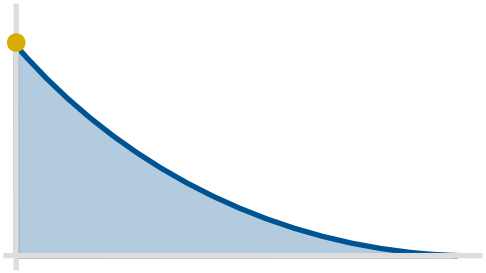
$$\rho(z, X) = \sum_{x \in X} \zeta^t(x, z)\mathbf{w}_X^t(x)$$

Defined for any point z. It equals 1 at every point of X.

Background

Magnitude Potential

SIMILARITY



$$\zeta^t(x, y) = exp(-t \cdot d(x, y))$$

The scale t sets how quickly similarity decays with distance d.

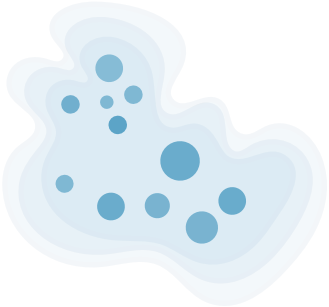
WEIGHTING



$$\mathbf{w}_X^t = (\zeta^t)^{-1}\mathbf{1}$$

The weighted similarities sum to 1 at every point of X.

MAGNITUDE POTENTIAL



$$\rho(z, X) = \sum_{x \in X} \zeta^t(x, z)\mathbf{w}_X^t(x)$$

Defined for any point z. It equals 1 at every point of X.

How to read ρ(z, X)

Coverage

A soft score of how well X covers z, decaying smoothly away from X.

Balanced Density

Exponential coverage, balanced by weights.

Smooth

The unique minimal-norm RKHS function equal to 1 on X (Meckes).

OUR FRAMEWORK

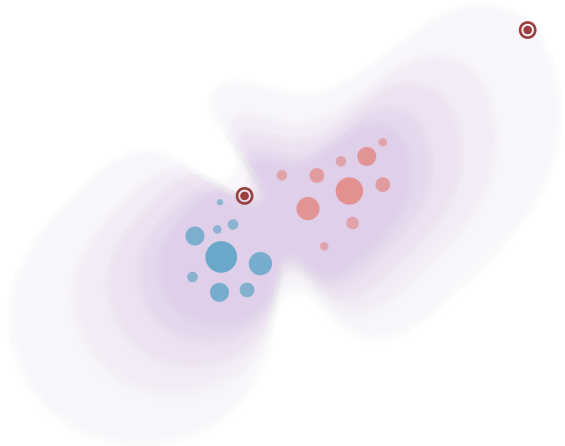
Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).

OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).

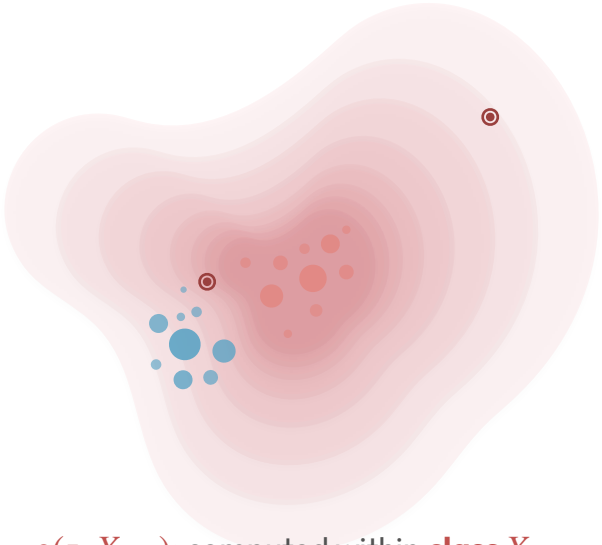


$\rho(z, X)$: computed within **full dataset** X

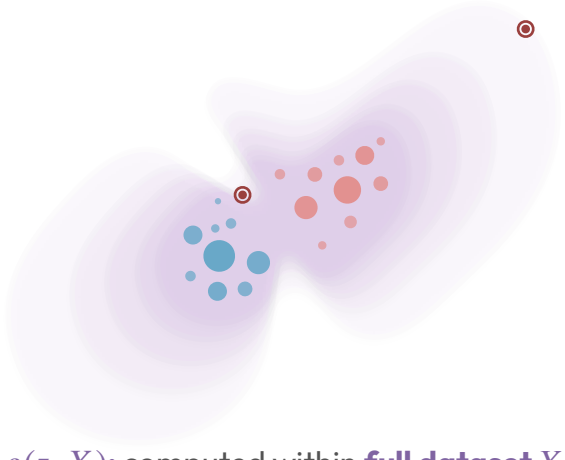
OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).



$\rho(z, X_{Red})$**:** computed within **class** X_{Red}

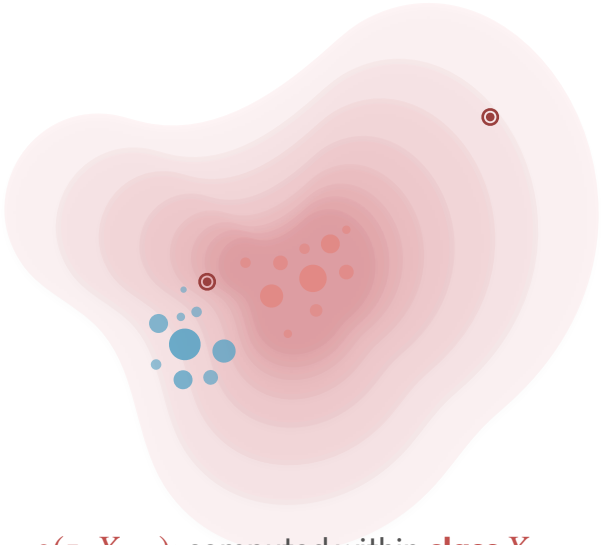


$\rho(z, X)$**:** computed within **full dataset** X

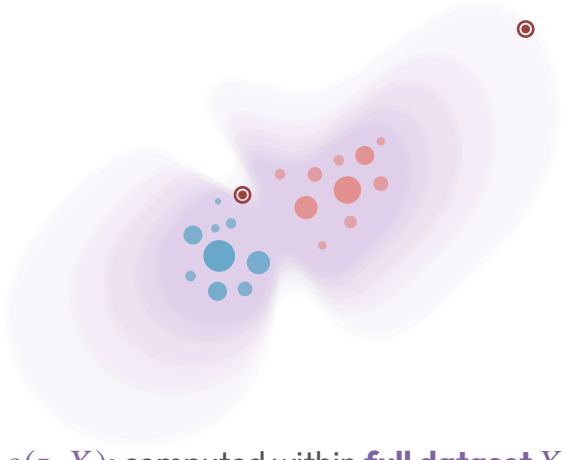
OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).



$\rho(z, X_{Red})$**:** computed within **class** X_{Red}



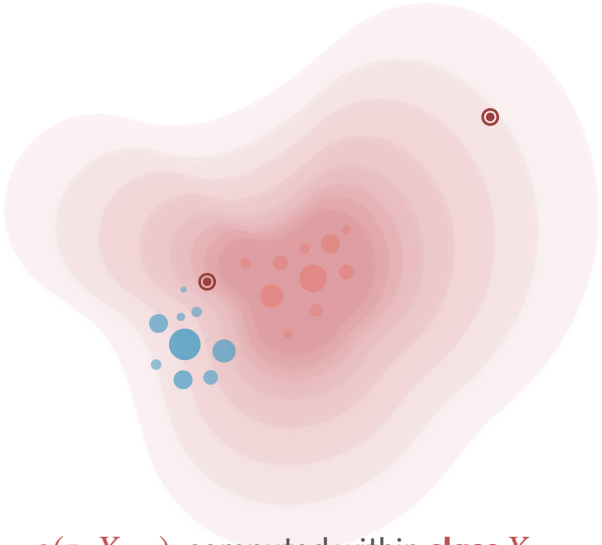
$\rho(z, X)$**:** computed within **full dataset** X

- **Magnitude potential ratio:** How well is z (◉) covered by its class X_c vs. the full dataset X?

OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).



$\rho(z, X_{Red})$**:** computed within **class** X_{Red}



$\rho(z, X)$**:** computed within **full dataset** X

- **Magnitude potential ratio:** How well is z (◉) covered by its **class** X_c vs. the full dataset X?

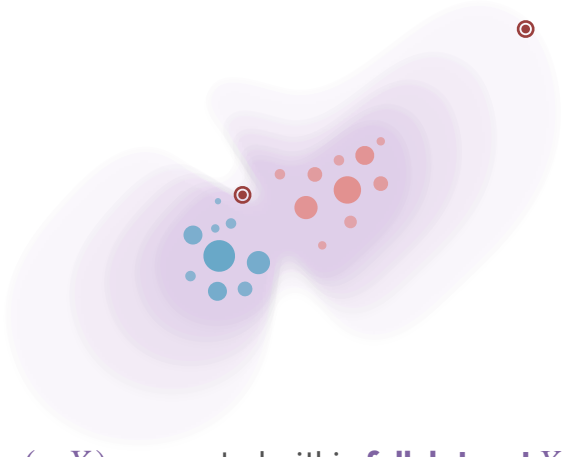
OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).



$\rho(z, X_{Red})$**:** computed within **class** X_{Red}



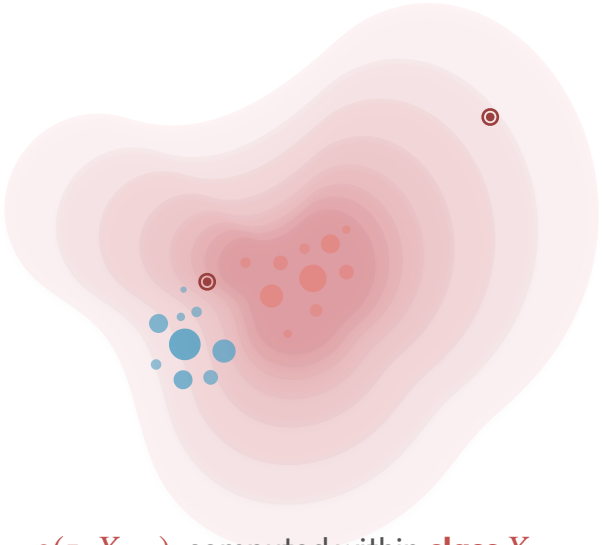
$\rho(z, X)$**:** computed within **full dataset** X

- **Magnitude potential ratio:** How well is z (◉) covered by its class X_c vs. the **full dataset** X?

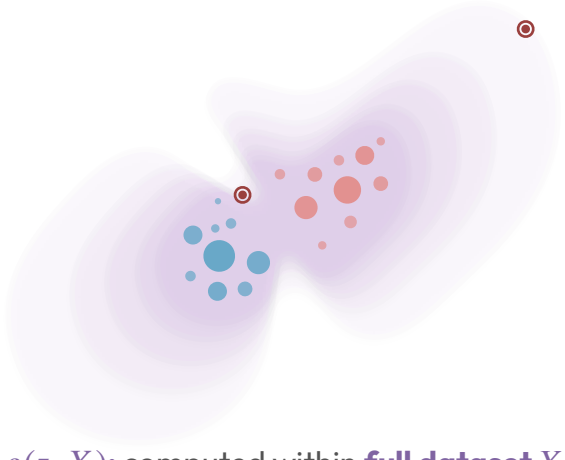
OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).



$\rho(z, X_{Red})$**:** computed within **class** X_{Red}



$\rho(z, X)$**:** computed within **full dataset** X

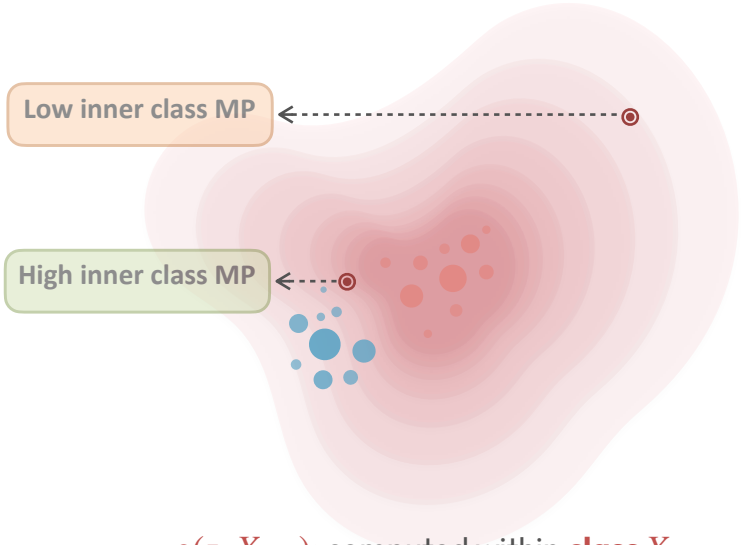
- **Magnitude potential ratio:** How well is z (◉) covered by its **class** X_c vs. the **full dataset** X?

$$R_{X,X_c}(z) = \rho(z, X_c)/\rho(z, X)$$

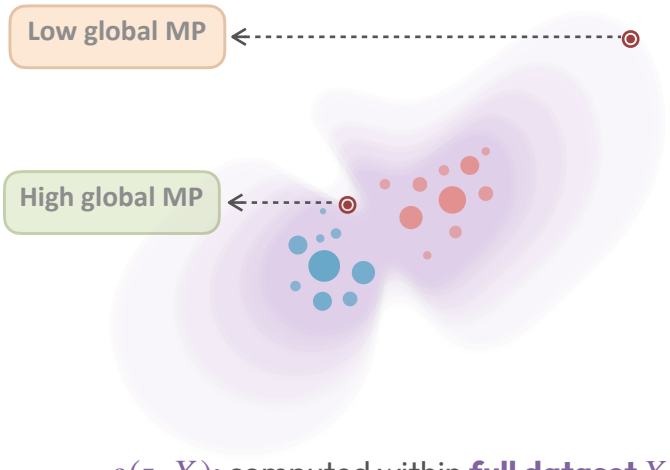
OUR FRAMEWORK

Magnitude Potential Ratio

- **Magnitude Potential $\rho(z, X)$:** Any dataset X generates a field — the total coverage felt at z (◉).



$\rho(z, X_{Red})$**:** computed within **class** X_{Red}



$\rho(z, X)$**:** computed within **full dataset** X

- **Magnitude potential ratio:** How well is z (◉) covered by its **class** X_c vs. the **full dataset** X?

$$R_{X,X_c}(z) = \rho(z, X_c)/\rho(z, X)$$

OUR FRAMEWORK

Magnitude Potential Ratio

How to use MP Ratio to answer our questions?

OUR FRAMEWORK

Magnitude Potential Ratio

How to use MP Ratio to answer our questions?

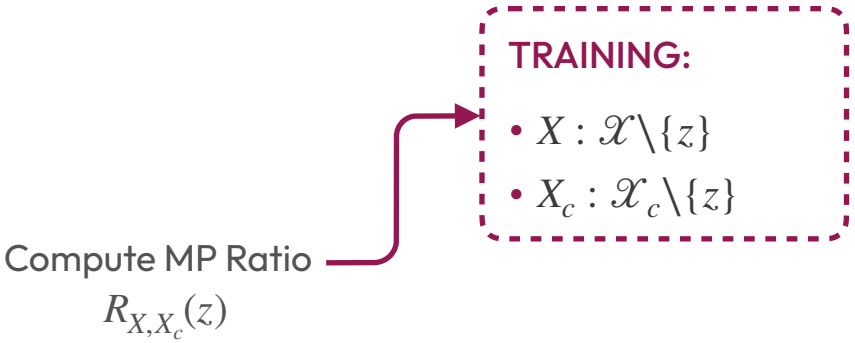
Compute MP Ratio

$$R_{X,X_c}(z)$$

OUR FRAMEWORK

Magnitude Potential Ratio

How to use MP Ratio to answer our questions?



$\mathcal{X}$ training data

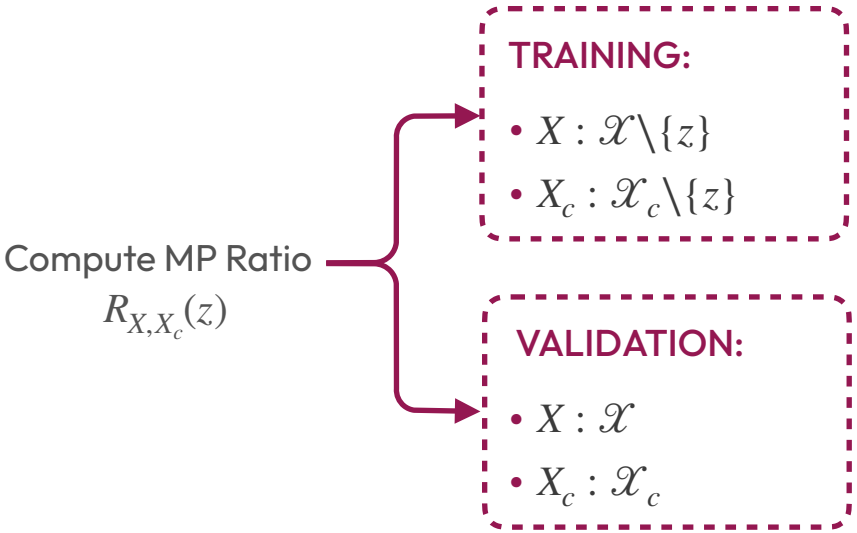
$\mathcal{X}_c$ training points in the class of z

- A training point z is in the data, where ρ is 1, so z is left out of both sets.

OUR FRAMEWORK

Magnitude Potential Ratio

How to use MP Ratio to answer our questions?



$\mathcal{X}$ training data

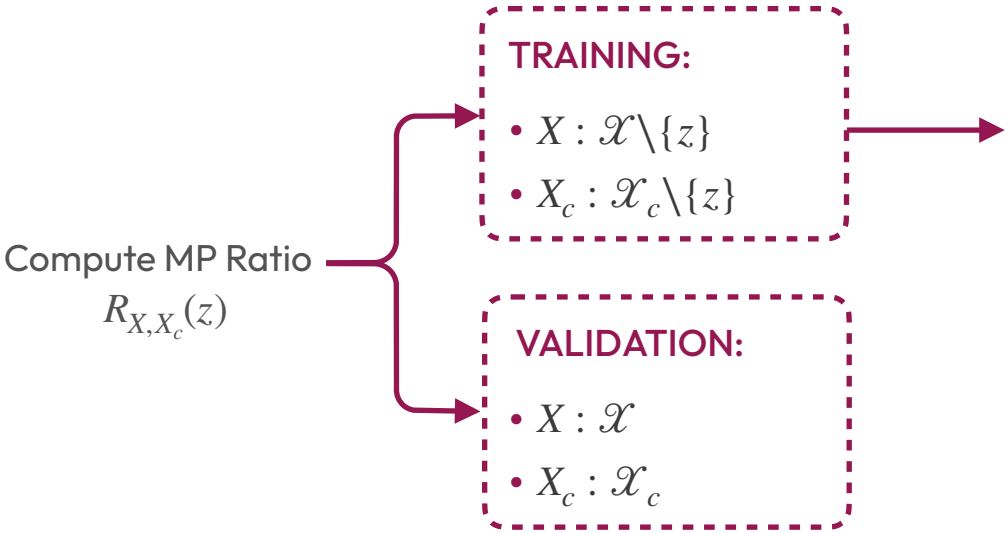
$\mathcal{X}_c$ training points in the class of z

- A training point z is in the data, where ρ is 1, so z is left out of both sets.
- A validation point is not in the data, so nothing is left out.

OUR FRAMEWORK

Magnitude Potential Ratio

How to use MP Ratio to answer our questions?



PER SAMPLE

Memorization

- Low R flags memorized points.
- R correlates negatively with Feldman scores on CIFAR-100.

$\mathcal{X}$ training data

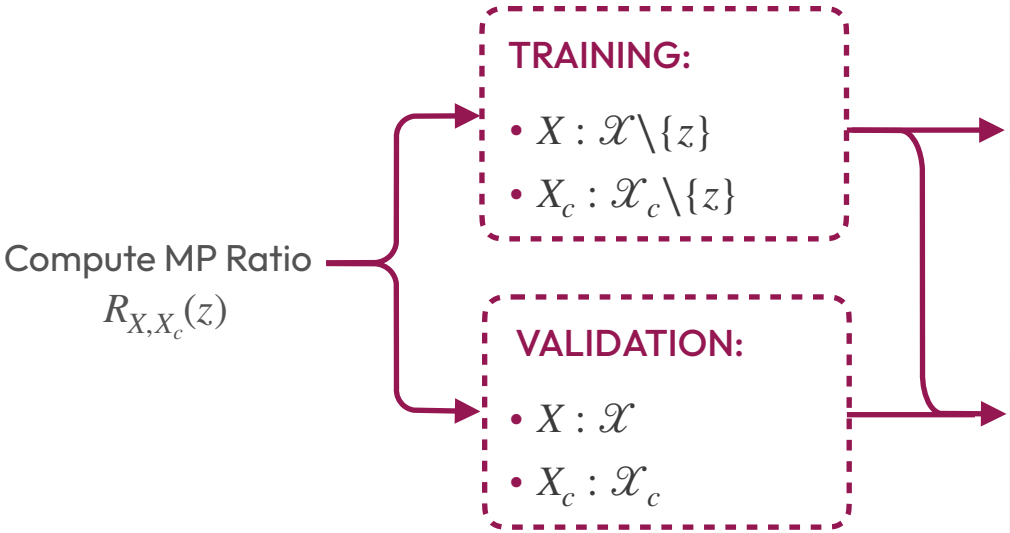
$\mathcal{X}_c$ training points in the class of z

- A training point z is in the data, where ρ is 1, so z is left out of both sets.
- A validation point is not in the data, so nothing is left out.

OUR FRAMEWORK

Magnitude Potential Ratio

How to use MP Ratio to answer our questions?



PER SAMPLE

Memorization

- Low R flags memorized points.
- R correlates negatively with Feldman scores on CIFAR-100.

AGGREGATION

Training dynamics

- Mean and standard deviation of R over epochs track training dynamics and decision boundaries.

$\mathcal{X}$ training data

$\mathcal{X}_c$ training points in the class of z

- A training point z is in the data, where ρ is 1, so z is left out of both sets.
- A validation point is not in the data, so nothing is left out.

RESULTS

Memorization on CIFAR-100

Measure	Spearman ρ	Top-500 overlap
Max softmax probability	-0.507	3.8% (19/500)
Logit margin	-0.655	3.6% (18/500)
Entropy	0.661	3.8% (19/500)
Cross-entropy loss	0.507	3.8% (19/500)
MP ratio, t = 0.05	-0.620	12.0% (60/500)
MP ratio, t = 0.10	-0.703	17.0% (85/500)
MP ratio, t = 0.15	-0.751	17.6% (88/500)
MP ratio, t = 0.20	-0.763	17.6% (88/500)
MP ratio, t = 0.25	-0.675	10.0% (50/500)

SETUP

ResNet-50 at epoch 160.

FELDMAN SCORES

Feldman scores from 4,000 models.

SIGN

Low MP ratio marks memorized points, so the correlation is negative.

TOP 500

At best the MP ratio recovers 88 of the 500 most memorized points. The baselines recover at most 19.

MP ratio rows average 10 random subsamples.

RESULTS

Memorization on CIFAR-100

Measure	Spearman ρ	Top-500 overlap
Max softmax probability	-0.507	3.8% (19/500)
Logit margin	-0.655	3.6% (18/500)
Entropy	0.661	3.8% (19/500)
Cross-entropy loss	0.507	3.8% (19/500)
MP ratio, t = 0.05	-0.620	12.0% (60/500)
MP ratio, t = 0.10	-0.703	17.0% (85/500)
MP ratio, t = 0.15	-0.751	17.6% (88/500)
MP ratio, t = 0.20	-0.763	17.6% (88/500)
MP ratio, t = 0.25	-0.675	10.0% (50/500)

SETUP

ResNet-50 at epoch 160.

FELDMAN SCORES

Feldman scores from **4,000** models.

SIGN

Low MP ratio marks memorized points, so the correlation is negative.

TOP 500

At best the MP ratio recovers 88 of the 500 most memorized points. The baselines recover at most 19.

MP ratio rows average 10 random subsamples.

RESULTS

Memorization on CIFAR-100

Measure	Spearman ρ	Top-500 overlap
Max softmax probability	-0.507	3.8% (19/500)
Logit margin	-0.655	3.6% (18/500)
Entropy	0.661	3.8% (19/500)
Cross-entropy loss	0.507	3.8% (19/500)
MP ratio, t = 0.05	**-0.620**	12.0% (60/500)
MP ratio, t = 0.10	**-0.703**	17.0% (85/500)
MP ratio, t = 0.15	**-0.751**	17.6% (88/500)
MP ratio, t = 0.20	**-0.763**	17.6% (88/500)
MP ratio, t = 0.25	**-0.675**	10.0% (50/500)

SETUP
ResNet-50 at epoch 160.

FELDMAN SCORES
Feldman scores from 4,000 models.

SIGN
Low MP ratio marks memorized points, so the correlation is negative.

TOP 500
At best the MP ratio recovers 88 of the 500 most memorized points. The baselines recover at most 19.

MP ratio rows average 10 random subsamples.

RESULTS

Memorization on CIFAR-100

Measure	Spearman ρ	Top-500 overlap
Max softmax probability	-0.507	3.8% (19/500)
Logit margin	-0.655	3.6% (18/500)
Entropy	0.661	3.8% (19/500)
Cross-entropy loss	0.507	3.8% (19/500)
MP ratio, t = 0.05	-0.620	**12.0% (60/500)**
MP ratio, t = 0.10	-0.703	**17.0% (85/500)**
MP ratio, t = 0.15	-0.751	**17.6% (88/500)**
MP ratio, t = 0.20	-0.763	**17.6% (88/500)**
MP ratio, t = 0.25	-0.675	**10.0% (50/500)**

SETUP
ResNet-50 at epoch 160.

FELDMAN SCORES
Feldman scores from 4,000 models.

SIGN
Low MP ratio marks memorized points, so the correlation is negative.

TOP 500
At best the MP ratio recovers 88 of the 500 most memorized points. The baselines recover at most 19.

MP ratio rows average 10 random subsamples.

RESULTS

Memorization on CIFAR-100

Measure	Spearman ρ	Top-500 overlap
Max softmax probability	-0.507	3.8% (19/500)
Logit margin	-0.655	3.6% (18/500)
Entropy	0.661	3.8% (19/500)
Cross-entropy loss	0.507	3.8% (19/500)
MP ratio, t = 0.05	-0.620	12.0% (60/500)
MP ratio, t = 0.10	-0.703	17.0% (85/500)
MP ratio, t = 0.15	-0.751	17.6% (88/500)
MP ratio, t = 0.20	-0.763	17.6% (88/500)
MP ratio, t = 0.25	-0.675	10.0% (50/500)

SETUP
ResNet-50 at epoch 160.

FELDMAN SCORES
Feldman scores from 4,000 models.

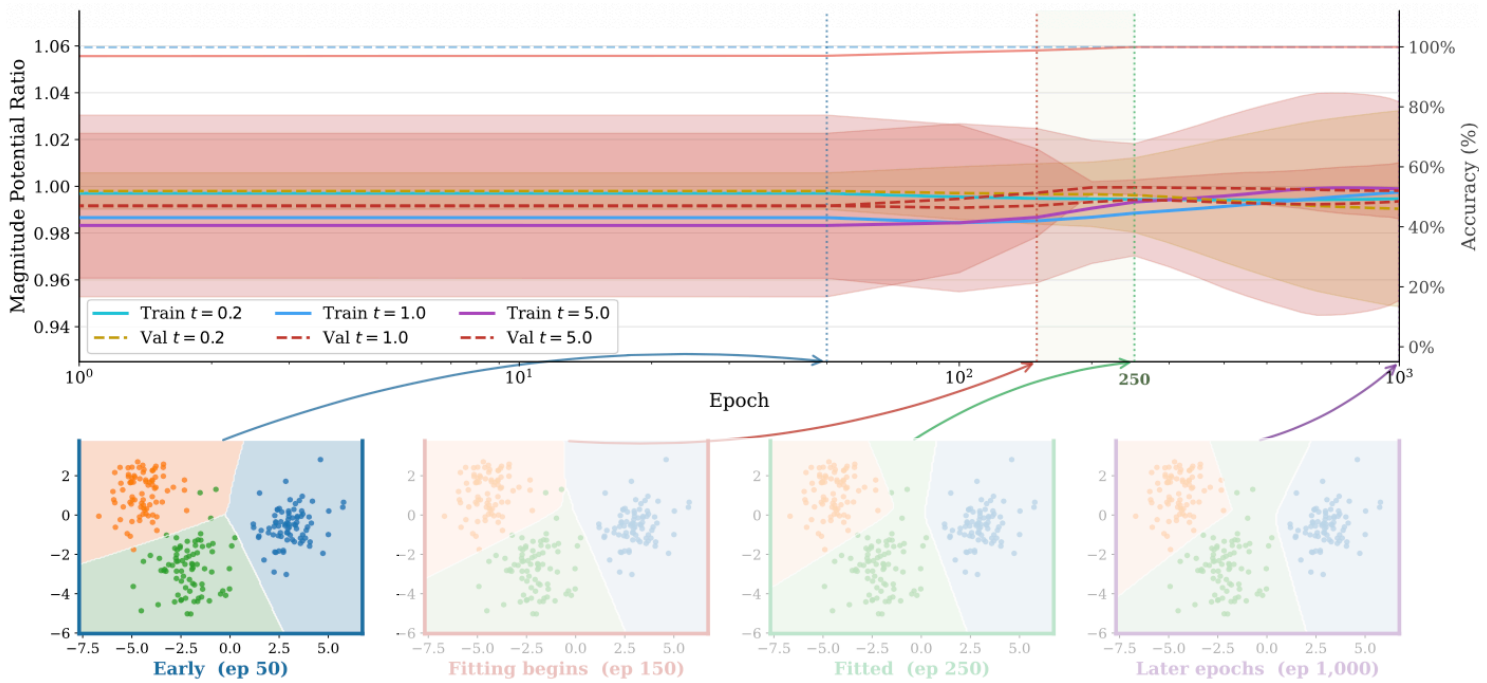
SIGN
Low MP ratio marks memorized points, so the correlation is negative.

TOP 500
At best the MP ratio recovers 88 of the 500 most memorized points. The baselines recover at most 19.

MP ratio rows average 10 random subsamples.

RESULTS

Decision Boundary and Overfitting



EARLY

Near-linear boundaries. The spread of the validation MP ratio is steady.

EPOCHS 150 TO 250

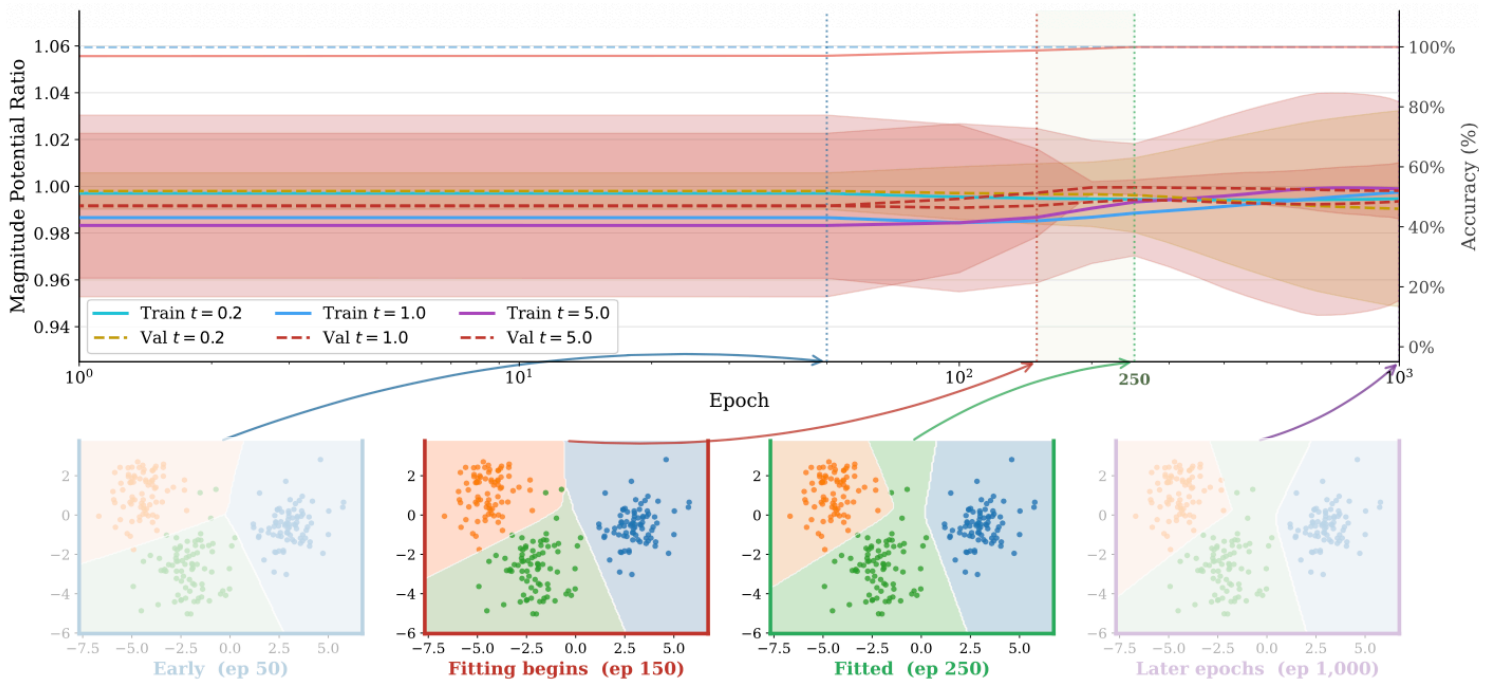
Boundaries become smooth, class-specific curves. The spread drops.

AFTER EPOCH 250

The spread rises again, a geometric signal of overfitting.

RESULTS

Decision Boundary and Overfitting



EARLY

Near-linear boundaries. The spread of the validation MP ratio is steady.

EPOCHS 150 TO 250

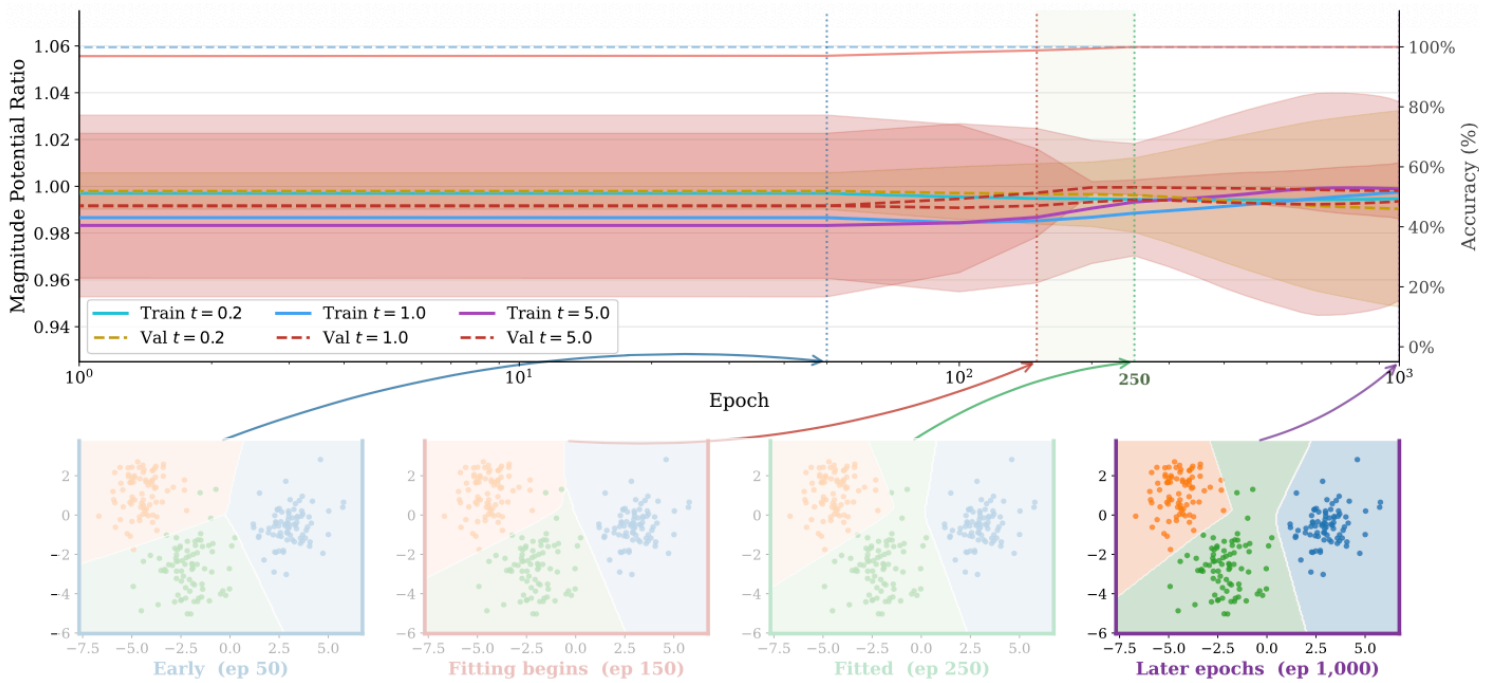
Boundaries become smooth, class-specific curves. The spread drops.

AFTER EPOCH 250

The spread rises again, a geometric signal of overfitting.

RESULTS

Decision Boundary and Overfitting



EARLY

Near-linear boundaries. The spread of the validation MP ratio is steady.

EPOCHS 150 TO 250

Boundaries become smooth, class-specific curves. The spread drops.

AFTER EPOCH 250

The spread rises again, a geometric signal of overfitting.

RESULTS

Grokking in Modular Addition

Grokking. Generalization arrives suddenly and long after the network has learned the training data.

RESULTS

Grokking in Modular Addition

Grokking. Generalization arrives suddenly and long after the network has learned the training data.

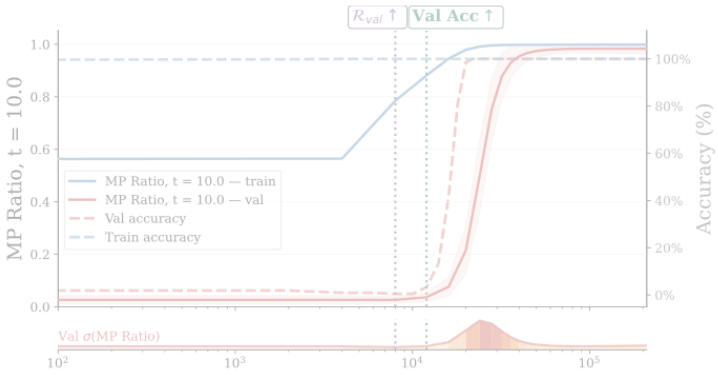
Why it matters. An early signal helps avoid terminating premature models that still have the potential to generalize.

RESULTS

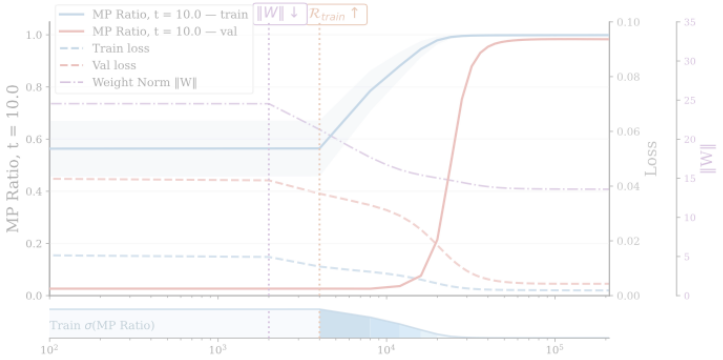
Grokking in Modular Addition

Grokking. Generalization arrives long after the network has memorized the training data.

Why it matters. An early signal helps avoid terminating premature models that still have the potential to generalize.



Against accuracy. The validation MP ratio and its std rise as validation accuracy begins its sharp rise.



Against loss and weight norm. The training MP ratio rises before validation accuracy. The weight norm falls slightly earlier, but does not indicate stability.

Setup. Modular addition mod 37, 2-layer MLP, MSE loss, t = 10.

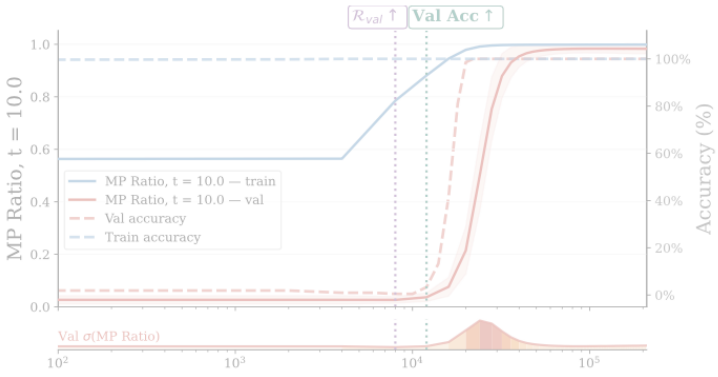
A 1-layer Transformer with cross-entropy shows the same dynamics.

RESULTS

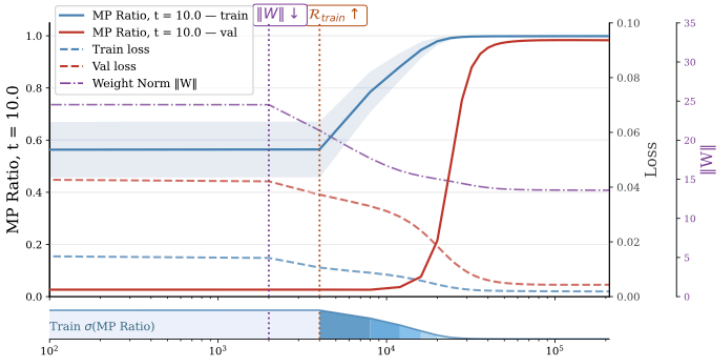
Grokking in Modular Addition

Grokking. Generalization arrives suddenly and long after the network has learned the training data.

Why it matters. An early signal helps avoid terminating premature models that still have the potential to generalize.



Against accuracy. The validation MP ratio and its std rise as validation accuracy begins its sharp rise.



Against loss and weight norm. The training MP ratio rises before validation accuracy. The weight norm falls slightly earlier, but does not indicate stability.

Setup. Modular addition mod 37, 2-layer MLP, MSE loss, t = 10.

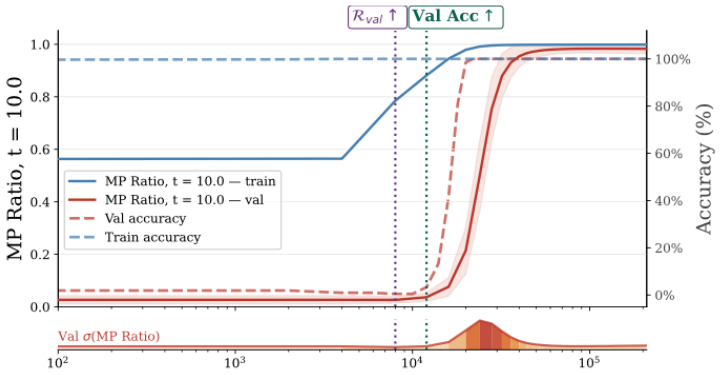
A 1-layer Transformer with cross-entropy shows the same dynamics.

RESULTS

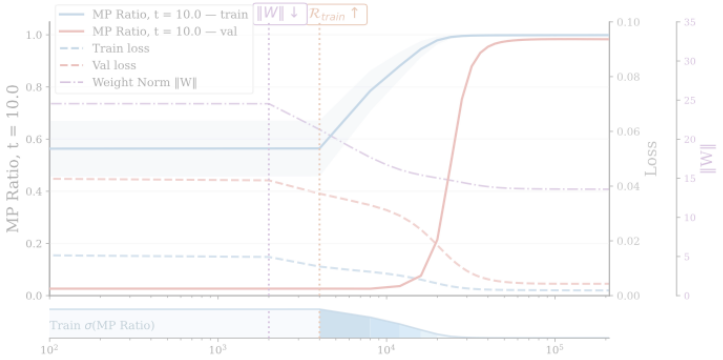
Grokking in Modular Addition

Grokking. Generalization arrives suddenly and long after the network has learned the training data.

Why it matters. An early signal helps avoid terminating premature models that still have the potential to generalize.



Against accuracy. The validation MP ratio and its std rise as validation accuracy begins its sharp rise.



Against loss and weight norm. The training MP ratio rises before validation accuracy. The weight norm falls slightly earlier, but does not indicate stability.

Setup. Modular addition mod 37, 2-layer MLP, MSE loss, t = 10.

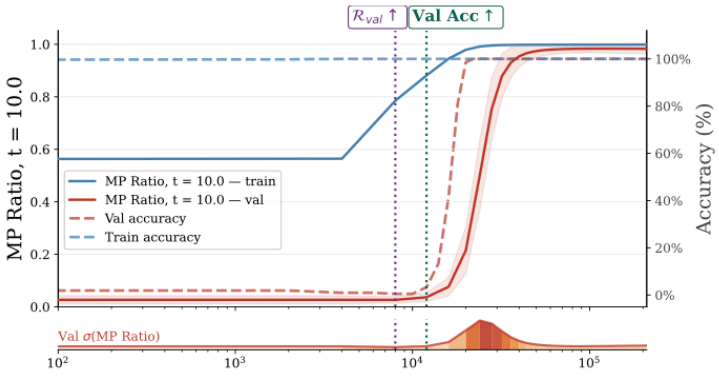
A 1-layer Transformer with cross-entropy shows the same dynamics.

RESULTS

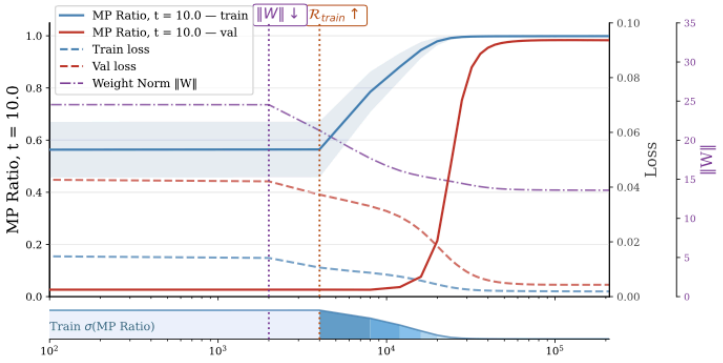
Grokking in Modular Addition

Grokking. Generalization arrives suddenly and long after the network has learned the training data.

Why it matters. An early signal helps avoid terminating premature models that still have the potential to generalize.



Against accuracy. The validation MP ratio and its std rise as validation accuracy begins its sharp rise.



Against loss and weight norm. The training MP ratio rises before validation accuracy. The weight norm falls slightly earlier, but does not indicate stability.

Setup. Modular addition mod 37, 2-layer MLP, MSE loss, t = 10.

A 1-layer Transformer with cross-entropy shows the same dynamics.

RESULTS

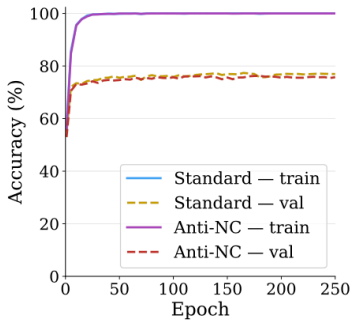
Theory

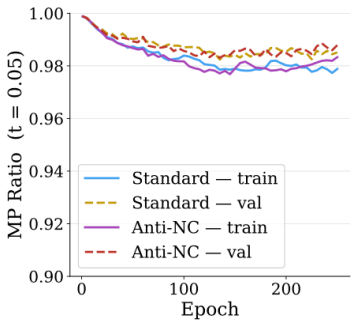
LIPSCHITZ	Theorem 3	Changes at most linearly with distance, so it is well behaved at unseen points.
SCALE LIMITS	Theorem 4	Small t: ρ tends to 1 everywhere. Large t: ρ is set by the nearest neighbour of z.
COLLAPSE	Corollary 5	When within-set variance goes to 0, the leave-one-out potential tends to 1.

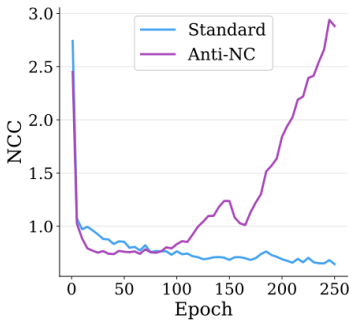
RESULTS

Theory

LIPSCHITZ	Theorem 3	Changes at most linearly with distance, so it is well behaved at unseen points.
SCALE LIMITS	Theorem 4	Small t: ρ tends to 1 everywhere. Large t: ρ is set by the nearest neighbour of z.
COLLAPSE	Corollary 5	When within-set variance goes to 0, the leave-one-out potential tends to 1.







NEURAL COLLAPSE

The MP ratio captures generalization structure whether or not neural collapse occurs.