# Geometric Analysis of Neural Regression Collapse via Intrinsic Dimension

Zixuan Dong* · Bimarsha Adhikari* · George Andriopoulos* · Keith Ross

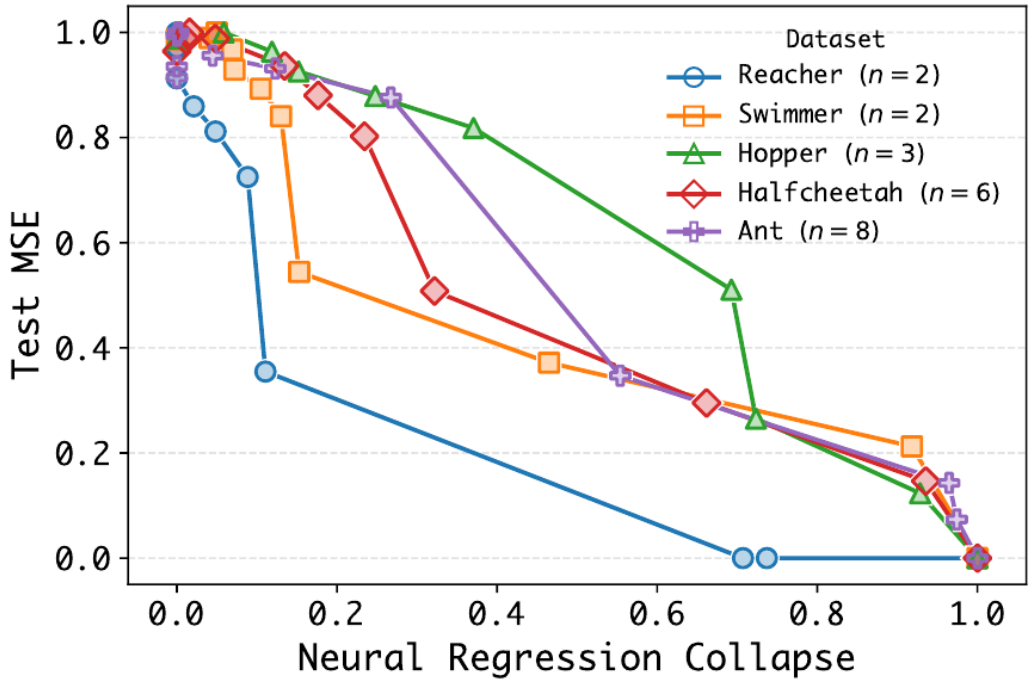
New York University · NYU Abu Dhabi · NYU Shanghai

NEURAL INFORMATION PROCESSING SYSTEMS

**NeurIPS 2026**

*Equal contribution

# Collapse helps classification but hurts regression




Five MuJoCo behavior-cloning datasets; smaller NRC = more collapsed.

## Regression heads are everywhere in control

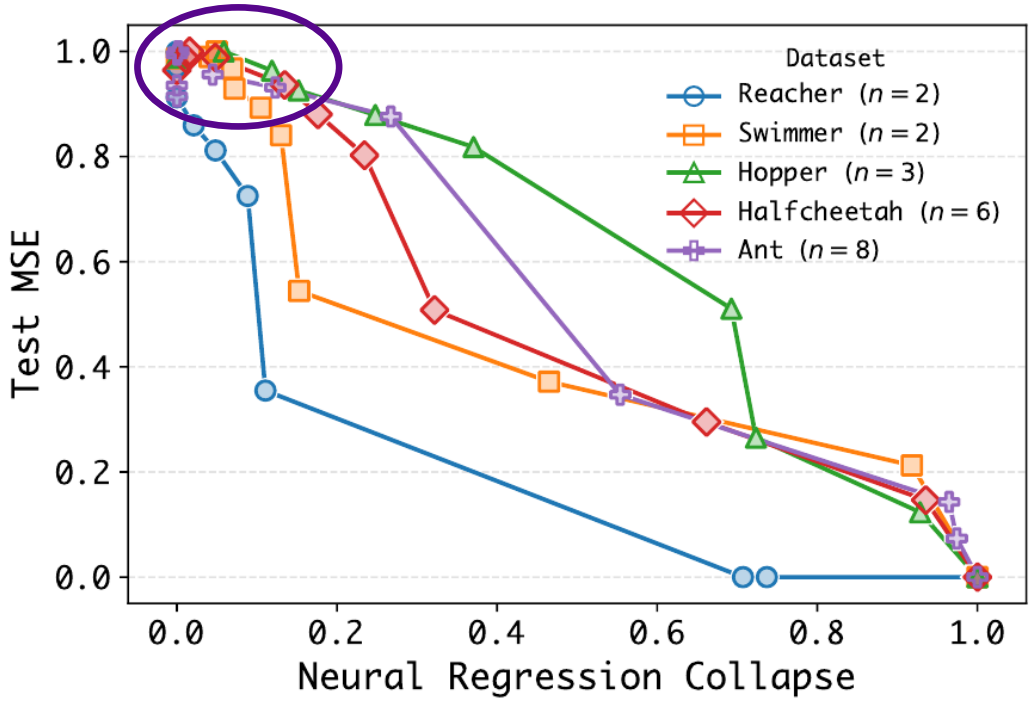
$$f(x) = W\,h_\theta(x) + b, \qquad W \in \mathbb{R}^{n\times d},\ \ n \ll d$$

Behavior cloning regresses actions; value critics regress Monte Carlo or bootstrapped returns.

## Neural regression collapse (NRC)

Last-layer features collapse onto the subspace of their top n principal components, n = number of targets. NRC1 → 0 means fully collapsed.

# Collapse helps classification but hurts regression




Five MuJoCo behavior-cloning datasets; smaller NRC = more collapsed.
**Circled: the most collapsed models have the highest test error.**

## Regression heads are everywhere in control

$$f(x) = W\,h_\theta(x) + b, \qquad W \in \mathbb{R}^{n\times d},\ \ n \ll d$$

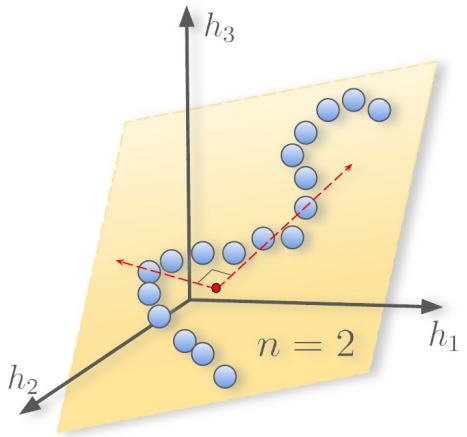
Behavior cloning regresses actions; value critics regress Monte Carlo or bootstrapped returns.

## Neural regression collapse (NRC)

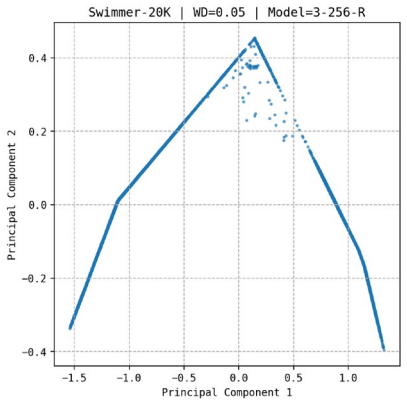
Last-layer features collapse onto the subspace of their top n principal components, n = number of targets. NRC1 → 0 means fully collapsed.

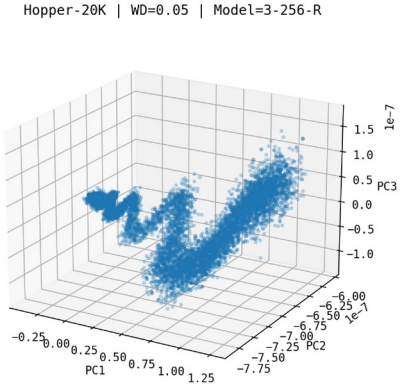
**In classification, collapse tends to help.**
**Why does it hurt in regression?**

# A linear lens misses curved features



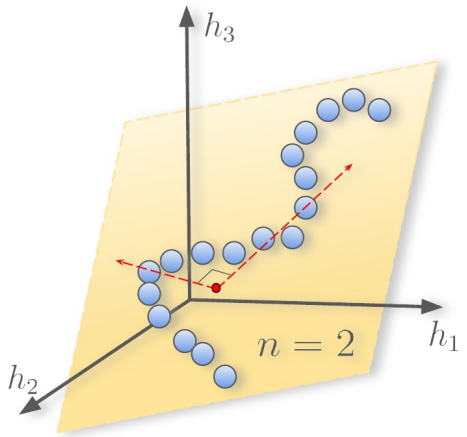

Features can curve inside the plane that PCA sees.



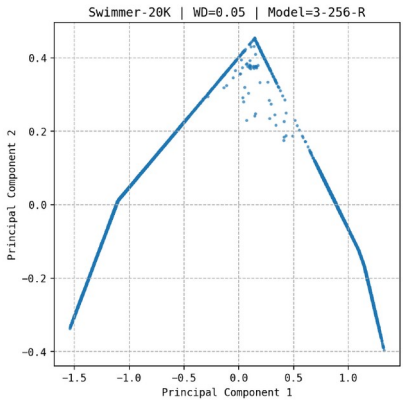



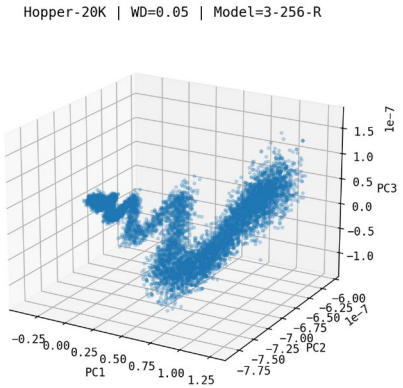

Collapsed Swimmer and Hopper features: 1-D curves.

# A linear lens misses curved features

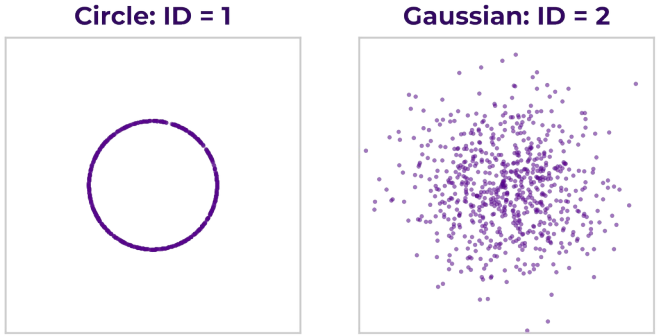



Features can curve inside the plane that PCA sees.



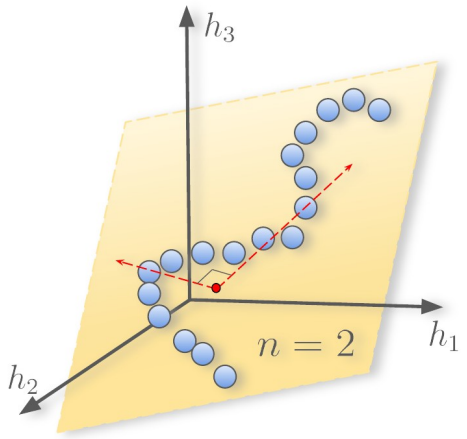




Collapsed Swimmer and Hopper features: 1-D curves.

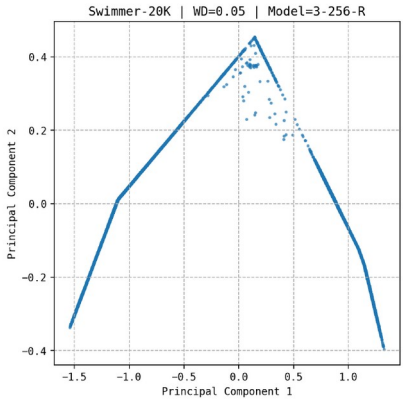


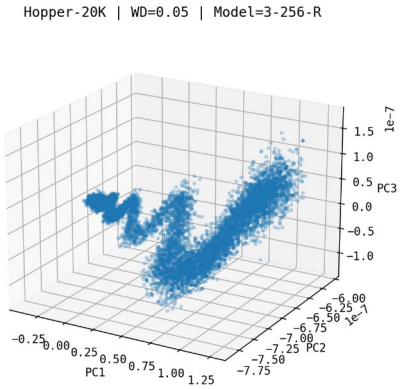

Same covariance, so the same NRC1 or effective rank.

# A linear lens misses curved features

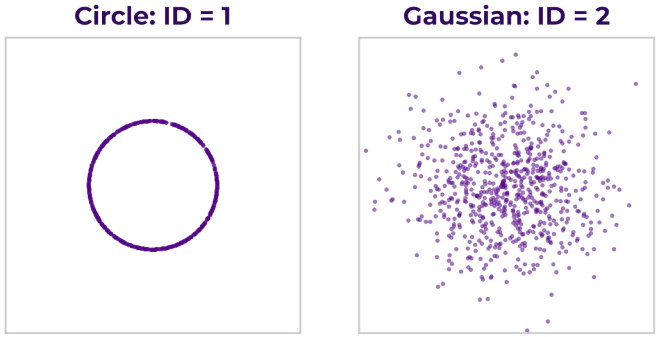



Features can curve inside the plane that PCA sees.




Collapsed Swimmer and Hopper features: 1-D curves.



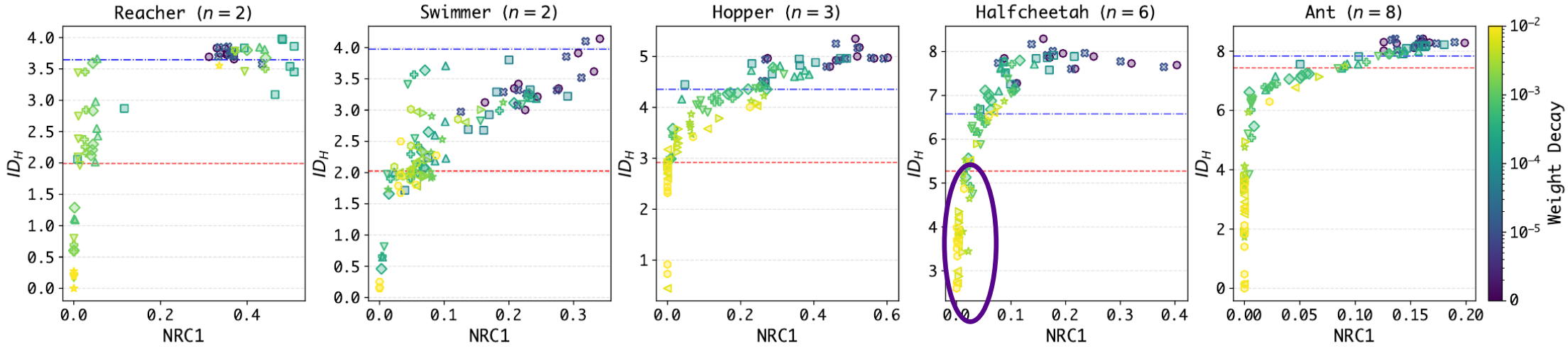

Same covariance, so the same NRC1 or effective rank.

## Intrinsic dimension (ID): the manifold's degrees of freedom

2-NN estimator: hyperparameter-free.
Compare the features' $ID_H$ with the targets' $ID_Y$.

$$\mu_i = \frac{r_2(x_i)}{r_1(x_i)}, \qquad \Pr(\mu > t) = t^{-d}$$

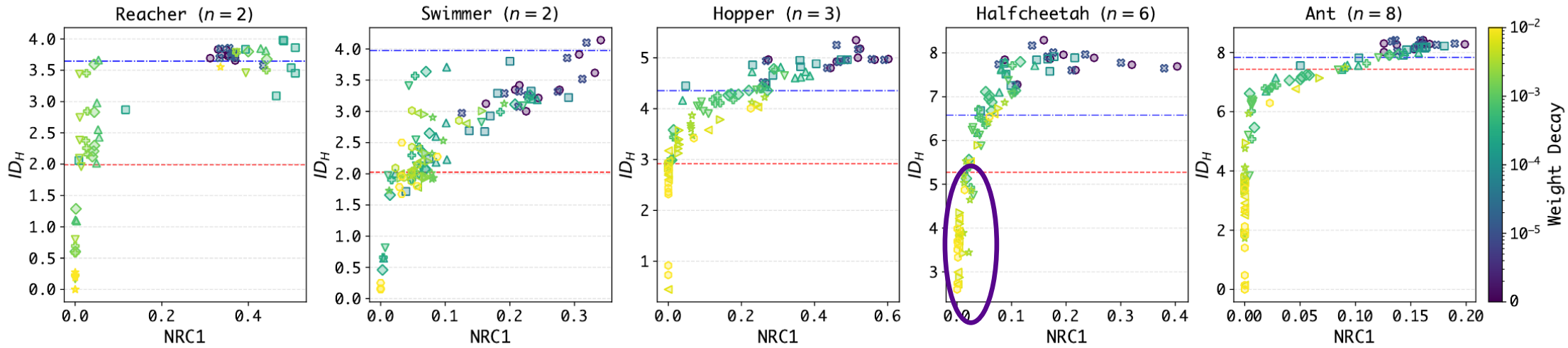
# Intrinsic dimension sees collapse that NRC1 misses




$ID_Y$ (targets)  $ID_X$ (inputs)

Each dot is a model; color = weight decay. Circled: collapsed Halfcheetah models.

# Intrinsic dimension sees collapse that NRC1 misses




$ID_Y$ (targets) $ID_X$ (inputs)

Each dot is a model; color = weight decay. Circled: collapsed Halfcheetah models.

**Collapsed: NRC1 ≈ 0**

$ID_H$ drops below $ID_Y$ and keeps shrinking, which NRC1 cannot see.

**Not collapsed**

$ID_H$ stays above $ID_Y$ and rises with NRC1. $ID_Y$ acts as a threshold.

# Three regimes, set by the target's ID and the data

### $ID_H < ID_Y$

**Over-compressed**

Underfits: high train and test error. This is why collapse hurts.

### $ID_H \approx ID_Y$

**Balanced**

Sweet spot when data is scarce or targets are noisy.

### $ID_H > ID_Y$

**Under-compressed**

Best with plenty of clean data: more dimension keeps helping.

# Three regimes, set by the target's ID and the data

### $ID_H < ID_Y$

**Over-compressed**

Underfits: high train and test error. This is why collapse hurts.

### $ID_H \approx ID_Y$

**Balanced**

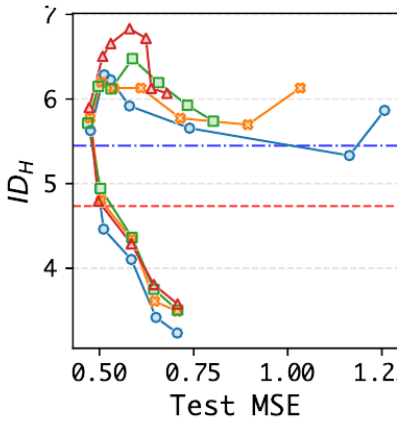
Sweet spot when data is scarce or targets are noisy.

### $ID_H > ID_Y$

**Under-compressed**

Best with plenty of clean data: more dimension keeps helping.

**Halfcheetah, 1K samples**




## Same task, two data sizes

**1K samples:** test error is U-shaped, lowest near $ID_Y$ (red dashed line).

Red dashed: $ID_Y$. Blue dash-dot: $ID_X$. Colors: model width.

# Three regimes, set by the target's ID and the data

### $ID_H < ID_Y$

**Over-compressed**

Underfits: high train and test error. This is why collapse hurts.

### $ID_H \approx ID_Y$

**Balanced**

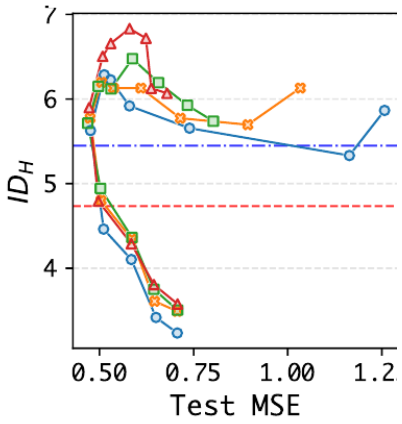
Sweet spot when data is scarce or targets are noisy.

### $ID_H > ID_Y$

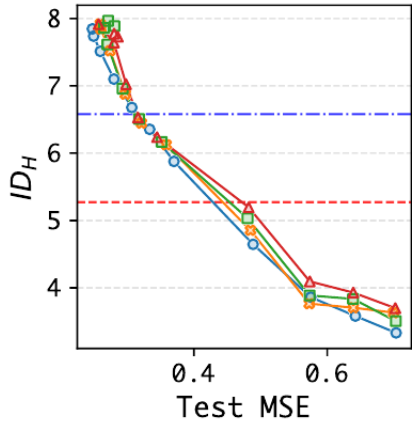
**Under-compressed**

Best with plenty of clean data: more dimension keeps helping.

**Halfcheetah, 1K samples**




**Halfcheetah, 20K samples**




## Same task, two data sizes

**1K samples:** test error is U-shaped, lowest near $ID_Y$ (red dashed line).

**20K samples:** test error keeps falling as $ID_H$ grows.

The same shift appears on Humanoid from 1K to 500K samples, and noisy targets (CIFAR-10) also give the U-shape.

Red dashed: $ID_Y$. Blue dash-dot: $ID_X$. Colors: model width.

# Why collapse happens, and why it hurts

## Weight decay selects the collapsed solution

Unconstrained feature model

$$\min_{W,H} \frac{1}{2M}\|WH - \tilde{Y}\|_F^2 + \frac{\lambda_W}{2}\|W\|_F^2 + \frac{\lambda_H}{2}\|H\|_F^2$$

$$\lambda = 0: \quad H = W^\dagger \tilde{Y} + (I_d - W^\dagger W)Z \ \text{ for any } Z$$

$$\lambda \to 0^+: \quad H_0 = W_0^\dagger \tilde{Y} \ \ (Z = 0)$$

Even vanishing regularization removes the null-space part: features lie in the n-dim row space of W.

$$c = \lambda_W \lambda_H \ \geq \ \lambda_n(\Sigma) \ \Rightarrow \ \operatorname{rank}(H) < n$$

Once c passes the smallest eigenvalue of the target covariance, rank drops and $ID_H$ can fall below $ID_Y$.

# Why collapse happens, and why it hurts

## Weight decay selects the collapsed solution

Unconstrained feature model

$$\min_{W,H} \frac{1}{2M}\|WH - \tilde{Y}\|_F^2 + \frac{\lambda_W}{2}\|W\|_F^2 + \frac{\lambda_H}{2}\|H\|_F^2$$

$$\lambda = 0: \quad H = W^\dagger \tilde{Y} + (I_d - W^\dagger W)Z \ \text{ for any } Z$$

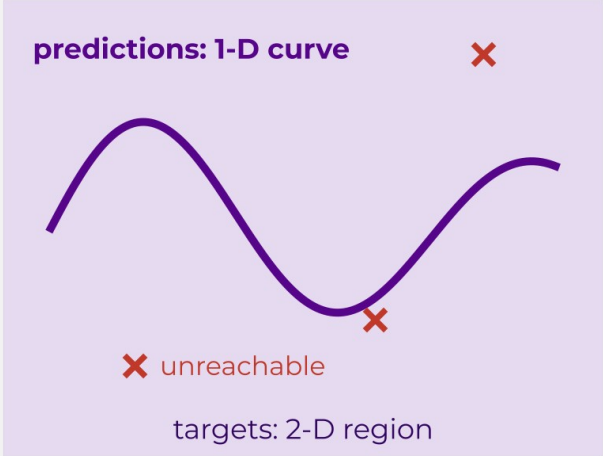
$$\lambda \to 0^+: \quad H_0 = W_0^\dagger \tilde{Y} \ \ (Z = 0)$$

Even vanishing regularization removes the null-space part: features lie in the n-dim row space of W.

$$c = \lambda_W \lambda_H \ \geq \ \lambda_n(\Sigma) \ \Rightarrow \ \text{rank}(H) < n$$

Once c passes the smallest eigenvalue of the target covariance, rank drops and $ID_H$ can fall below $ID_Y$.

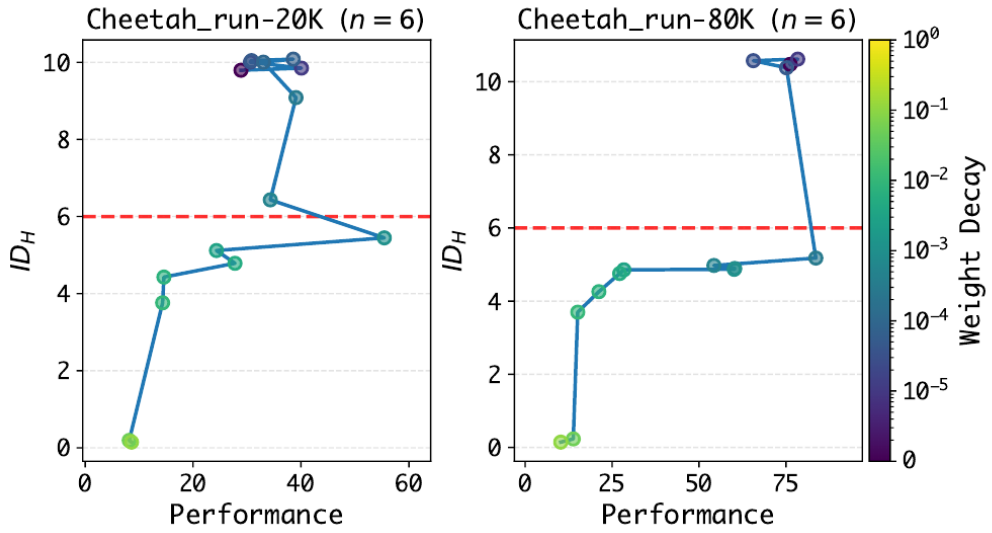
## Some targets become unreachable




$$m < n \ \Rightarrow \ g(\mathcal{M}_H) \text{ has measure zero in } \mathcal{N}_Y$$

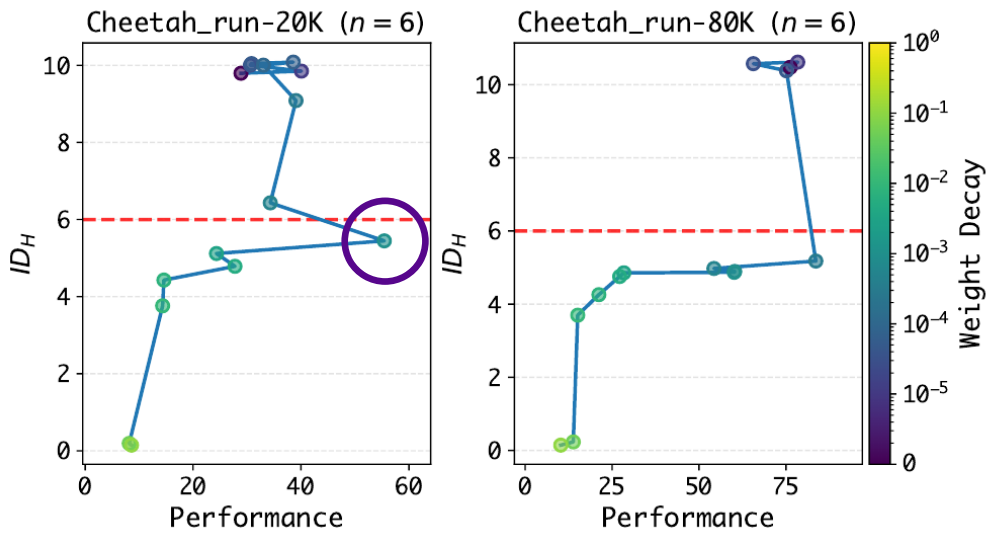
Sard's theorem: a smooth map from an m-dim feature manifold covers a measure-zero part of an n-dim target manifold.

# A rollout-free signal for choosing policies




Visual behavior cloning on Cheetah-run (pixels). Red dashed: $ID_\gamma$.

# A rollout-free signal for choosing policies



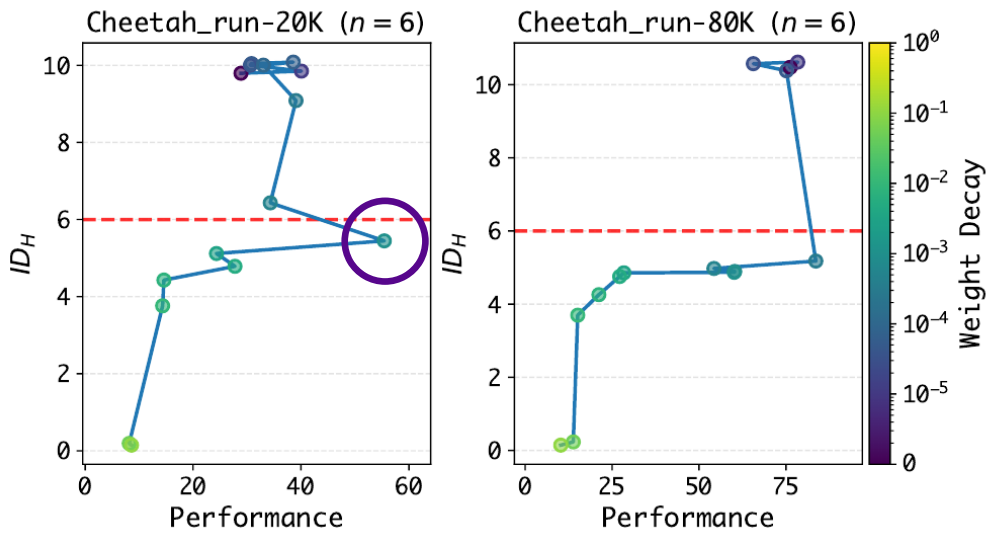

Visual behavior cloning on Cheetah-run (pixels). Red dashed: $ID_\gamma$.
**Circled: the best 20K policy, just below $ID_\gamma$.**

| Weight decay | 0 | 1e-4 | 5e-4 | 1e-3 | 5e-3 | 1e-2 |
|---|---|---|---|---|---|---|
| **NRC1** | 0.645 | 0.648 | 0.021 | **0.023** | 0.014 | 0.007 |
| **RankMe** | 73.1 | 69.9 | 3.54 | **3.63** | 3.57 | 3.16 |
| **$ID_H$** | 9.79 | 10.00 | 6.43 | **5.45** | 4.79 | 3.76 |
| **Test MSE** | 3.35 | 3.00 | 2.38 | **2.21** | 2.44 | 2.73 |
| **Return** | 28.9 | 33.0 | 34.3 | **55.4** | 27.8 | 14.4 |

20K demonstrations; $ID_\gamma$ = 6.14. Highlighted: best test MSE and return.

Weight decay ≥ 5e-4: NRC1 and RankMe are flat, yet return spans 14.4 to 55.4. Test MSE moves ~10% while return halves.

# A rollout-free signal for choosing policies




Visual behavior cloning on Cheetah-run (pixels). Red dashed: $ID_Y$.
**Circled: the best 20K policy, just below $ID_Y$.**

| Weight decay | 0 | 1e-4 | 5e-4 | 1e-3 | 5e-3 | 1e-2 |
|---|---|---|---|---|---|---|
| **NRC1** | 0.645 | 0.648 | 0.021 | **0.023** | 0.014 | 0.007 |
| **RankMe** | 73.1 | 69.9 | 3.54 | **3.63** | 3.57 | 3.16 |
| **$ID_H$** | 9.79 | 10.00 | 6.43 | **5.45** | 4.79 | 3.76 |
| **Test MSE** | 3.35 | 3.00 | 2.38 | **2.21** | 2.44 | 2.73 |
| **Return** | 28.9 | 33.0 | 34.3 | **55.4** | 27.8 | 14.4 |

20K demonstrations; $ID_Y$ = 6.14. Highlighted: best test MSE and return.

Weight decay ≥ 5e-4: NRC1 and RankMe are flat, yet return spans 14.4 to 55.4. Test MSE moves ~10% while return halves.

**Regime unknown in advance? Keep two candidates:**

Preliminary: one simulated benchmark.

1. **Largest $ID_H$** for plenty of clean data
2. **Smallest gap $|ID_H - ID_Y|$** for scarce or noisy data

# What this could enable

## In practice

1. **Cheap model selection**

   ID needs only nearest-neighbor distances on features and targets: a check next to validation loss when rollouts are costly.

2. **Tune regularization per task**

   Weight decay, dropout and depth all move $ID_H$. On Reacher, +0.0002 weight decay can tip a model into over-compression.

3. **Check the rest of the recipe**

   Normalization, augmentation, optimizer, batch size and early stopping may shift $ID_H$ too.

# What this could enable

## In practice

1. **Cheap model selection**

   ID needs only nearest-neighbor distances on features and targets: a check next to validation loss when rollouts are costly.

2. **Tune regularization per task**

   Weight decay, dropout and depth all move $ID_H$. On Reacher, +0.0002 weight decay can tip a model into over-compression.

3. **Check the rest of the recipe**

   Normalization, augmentation, optimizer, batch size and early stopping may shift $ID_H$ too.

## In theory

1. **Beyond the unconstrained feature model**

   Deep, nonlinear feature models that explain where along depth ID drops.

2. **From NRC to ID to generalization**

   Bound $ID_H$ with NRC, which UFM theory already characterizes, then connect to existing ID-based generalization results.

3. **When does a small ID gap help?**

   Which data properties favor it: sample size, target noise, target-manifold geometry.

# Regularization can over-compress regression features

Intrinsic dimension, not NRC1, shows when: below $ID_Y$, error rises; above it, the amount and noise of the data decide.

Geometric Analysis of Neural Regression Collapse via Intrinsic Dimension

Zixuan Dong · Bimarsha Adhikari · George Andriopoulos · Keith Ross

## Thanks for watching