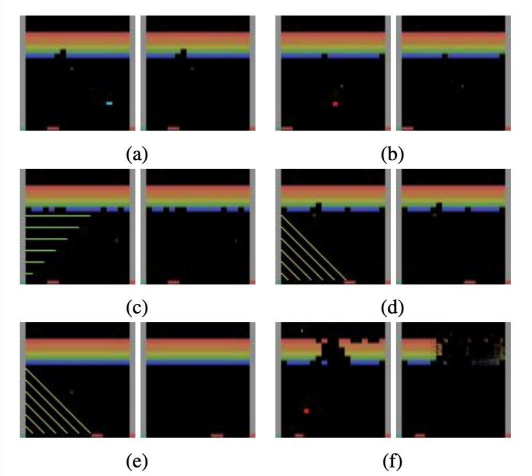
Observational Overfitting in Reinforcement Learning

Xingyou Song, Yiding Jiang, Stephen Tu, Yilun Du, Behnam Neyshabur

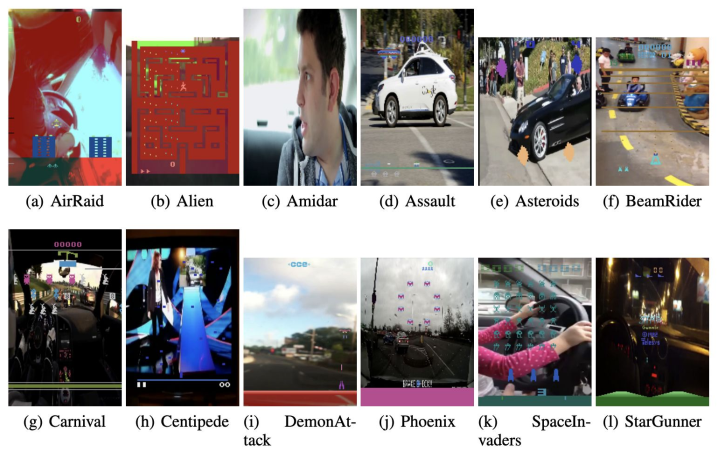
Google, MIT

Previous Works on RL Generalization

- Numerous Works investigating changing MDP backgrounds



[Gamrian18]

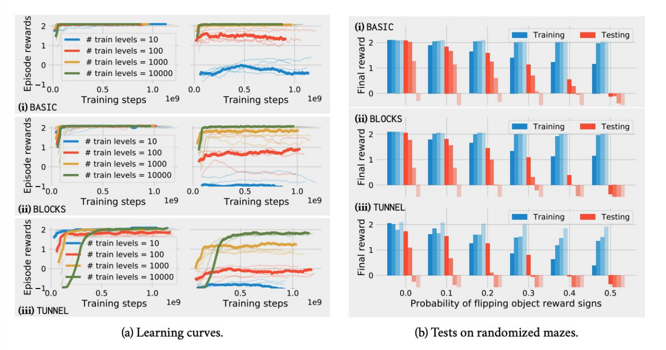


[Zhang18]

Previous Works on RL Generalization

- Other works showing that RL agents overfit, but not entirely from changing backgrounds:

[Cobbe18]



[Zhang17]

What does it mean to overfit in RL?

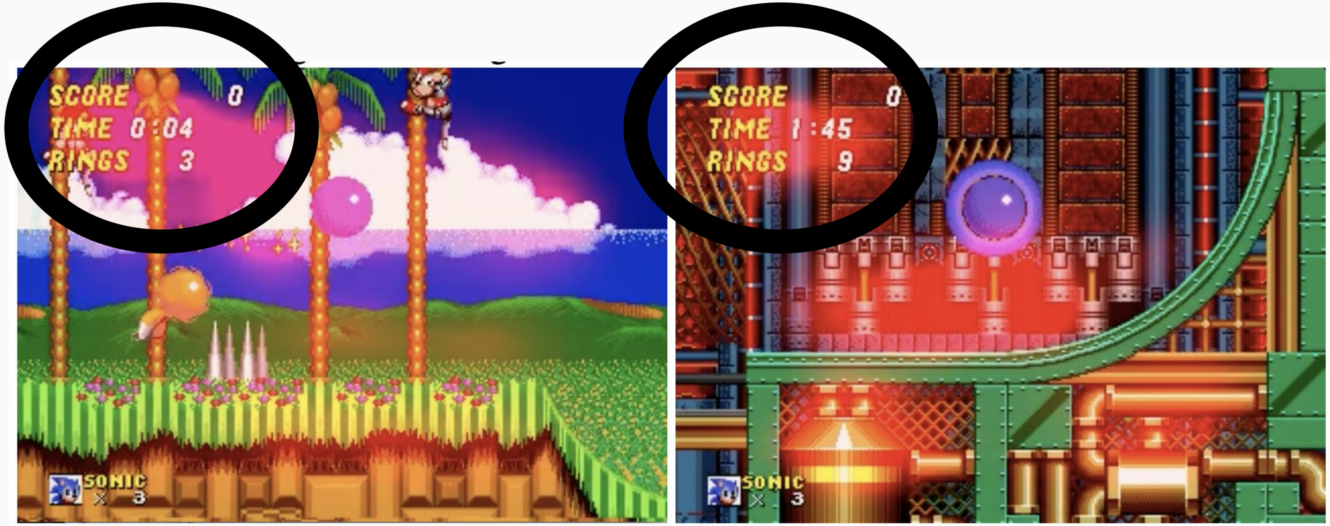
- Zero-Shot Generalization: Agent allowed finite training set of MDPs, evaluated on unseen test set of MDPs.
- Ideally, all MDP's sampled from a *distribution,* similar to Supervised Learning.
- Overfitting: Reward Gap between training and testing.

Current Work

- Sonic the HedgeHog (Gym Retro, [Nichol18]): Saliency (Red) suggests overfitting to background

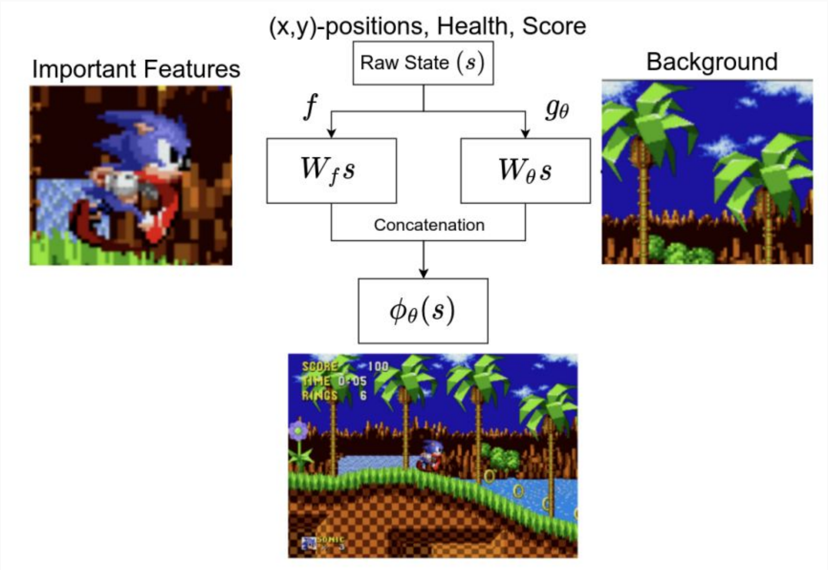
Current Work

- Agent can train even if it only saw the timer!



General Framework: "Observational Overfitting"

- For a fixed MDP $\mathcal{M}$, can generate multiple MDP's $\mathcal{M}_\theta$ by sampling "observation functions" $\phi_\theta : \mathcal{S} \rightarrow \mathcal{O}$
- Important invariant features projected from the same function f
- But background projection function g_θ changes per seed

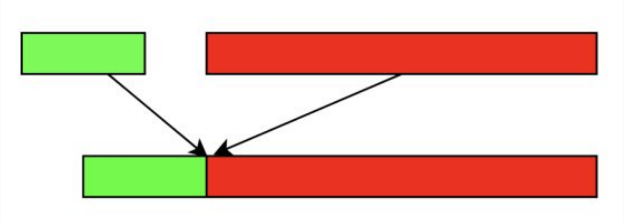


Base Case: LQR

- In the linear case, let $f(s) = W_f s$ and $g(s) = W_\theta s$
- A underlying cost $C(P)$ can be transformed into observation space cost

$$C(K; W_\theta) = C\left(K \begin{bmatrix} W_f \\ W_\theta \end{bmatrix}\right)$$

- If $P_\star$ is unique minimizer of $C(P)$, then multiple solutions $\begin{bmatrix} \alpha W_f P_\star^\intercal \\ (1-\alpha) W_\theta P_\star^\intercal \end{bmatrix}^\intercal$ are induced for $C(K; W_\theta)$; the only solution that generalizes is $\alpha = 1$



Theoretical Case: 1-Step LQR

- For a 1-Step LQR (convex) case, let

$$C(K; W_\theta) = \frac{1}{2}\left\| I + K \begin{bmatrix} W_f \\ W_\theta \end{bmatrix} \right\|_F^2$$

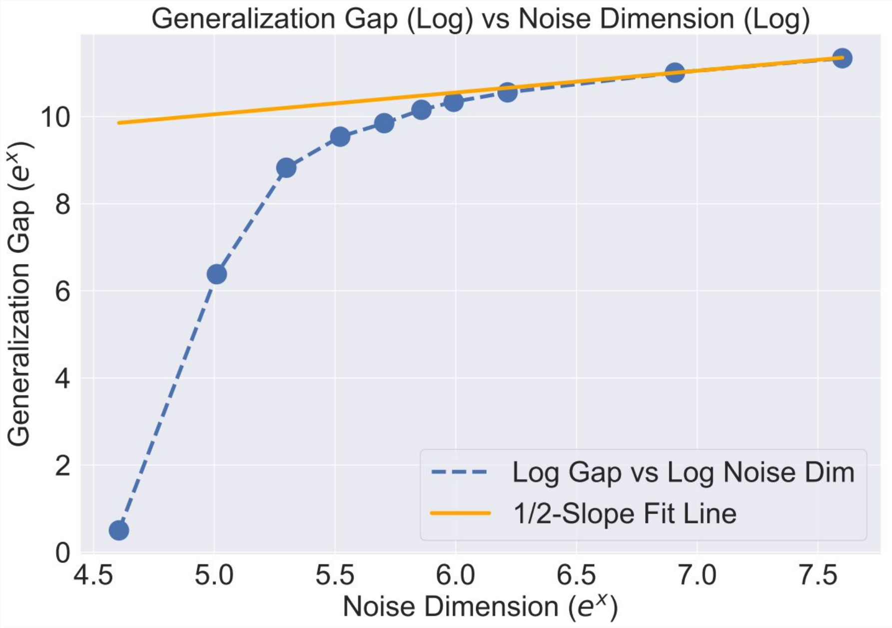
- Then

$$\nabla C(K; W_\theta) = \left(I + K \begin{bmatrix} W_f \\ W_\theta \end{bmatrix} \right) \begin{bmatrix} W_f \\ W_\theta \end{bmatrix}^\mathsf{T} \quad \nabla^2 C(K; W_\theta) = \begin{bmatrix} W_f \\ W_\theta \end{bmatrix} \begin{bmatrix} W_f \\ W_\theta \end{bmatrix}^\mathsf{T}$$

- Correct population minimizer lives in degenerate Hessian's span.
- Non-degenerate components of initialization do not change, hence overfitting must occur.

Experimental Case: Nonconvex LQR

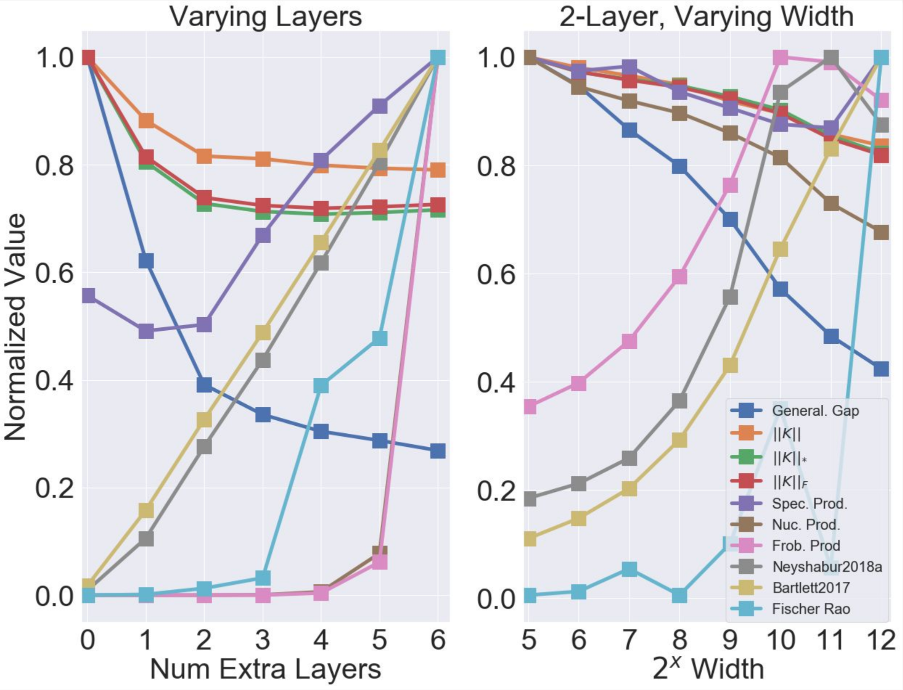
- For nonconvex full-LQR case, increasing g_θ dimension *increases* overfitting.
- This doesn't happen in the 1-Step convex case.



Experimental Case: Nonconvex LQR

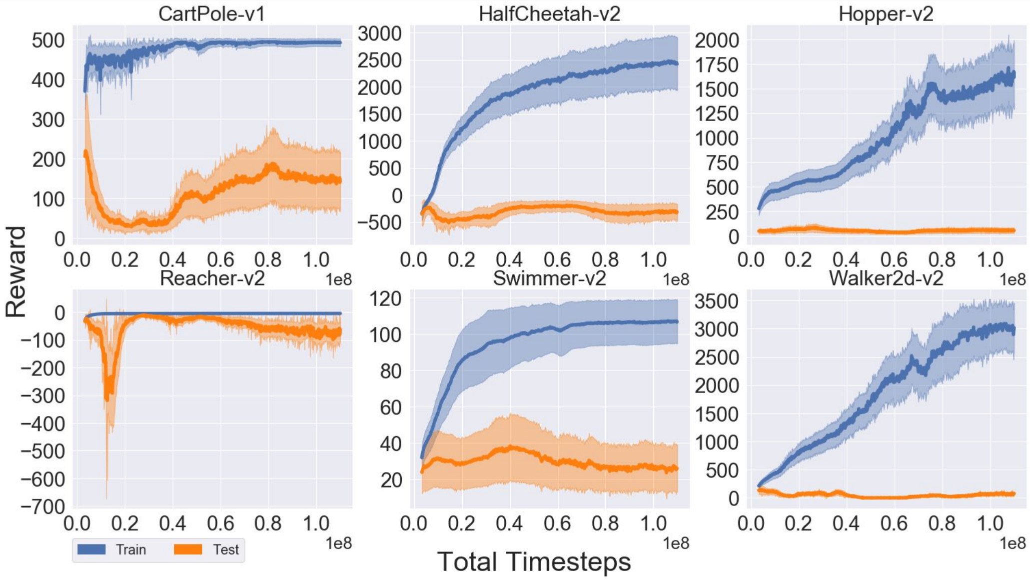
- Adding more linear layers reduces generalization gap in LQR.
- Many SL generalization bounds rely on using Lipschitz bounds, which LQR also satisfies.
- So can we upper bound the LQR generalization gap with SL bounds? Nope!
- Our theoretical understanding of RL generalization is limited.

$$K = K_1 K_2 \ldots K_j$$



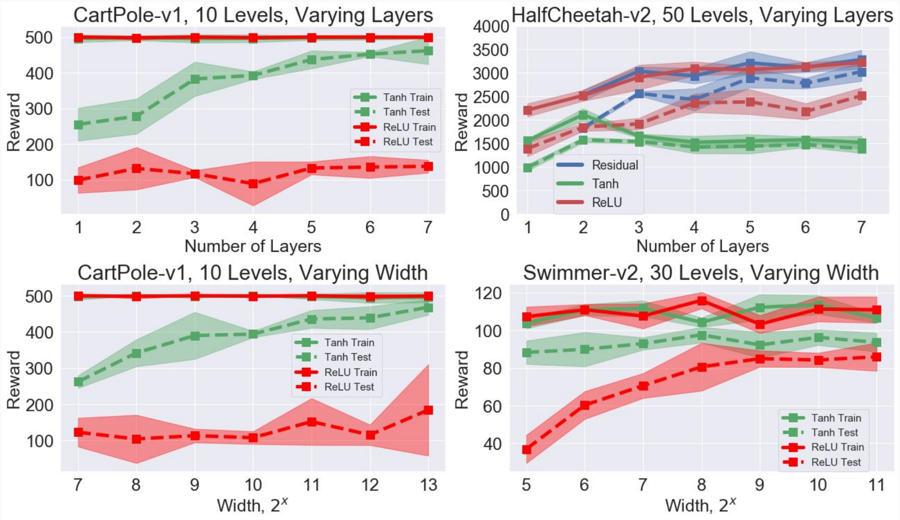
Nonlinear 1D Case

- Can we get a generalization gap using the same projection setup for Mujoco?
- Yes.
- Fixed number of levels for each environment, with same observation dimensions.



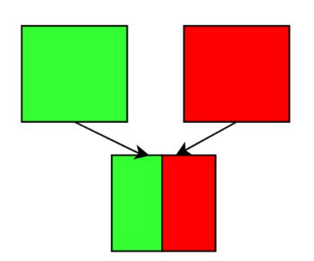
Nonlinear 1D Case

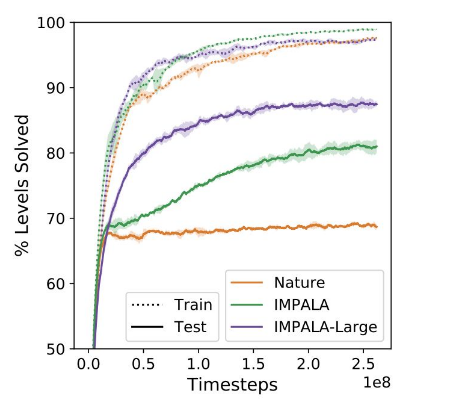
- Does overparameterization help?
- Yes! (But the effect can be dependent on choice of non-linearity.)
- Residual ReLU layers also improve generalization as well (HalfCheetah).



Nonlinear 2D (Image) Case

- What about 2-D case? We use linear deconvolutional layers to project a 1-D state to 2-D (84x84) classic DQN dimensions.
- We use the same architectures from CoinRun [Cobbe18], which increase generalization, in order: 1.NatureCNN, 2. IMPALA, 3. IMPALA-LARGE.



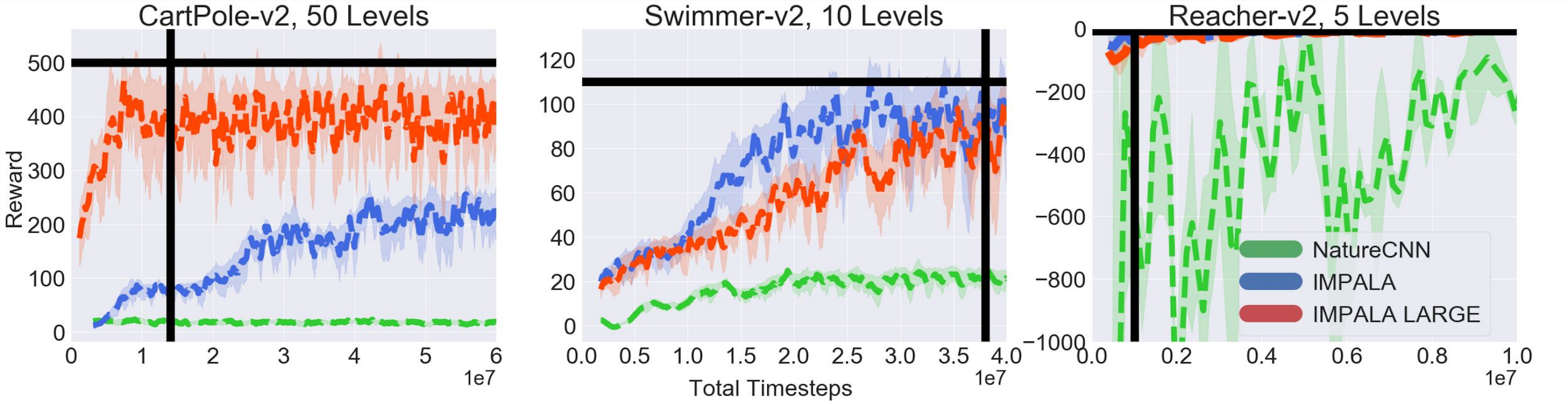


(b) Performance of Nature-CNN and IMPALA-CNN agents during training, on a set of 500 training levels.

[Cobbe18]

Nonlinear 2D (Image) Case

- Result: We get the ***same ranking under our projection case.***



Implicit Regularization in Reinforcement Learning

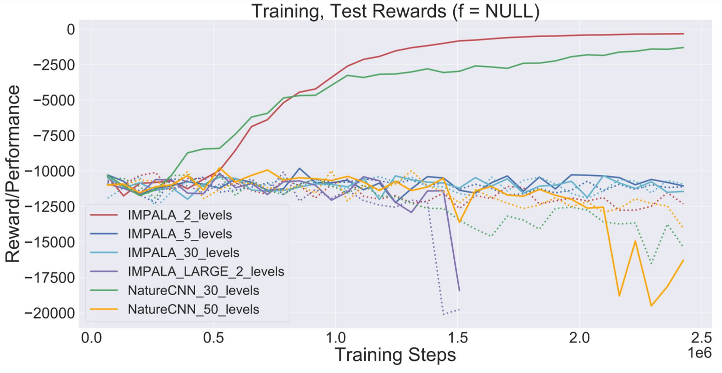
- How are all the above results related? ***Implicit Regularization.***
- “Implicit Regularization” [Neyshabur17]: any form of regularization not expressed in the end-to-end loss.
- Forms of implicit regularization in our work:
  - Overparameterization in neural network policies.
  - Special network modifications (Choice of non-linearity, Use of residual layers)
- Other forms from SL literature:
  - Choice of optimizer/Batch-Size.

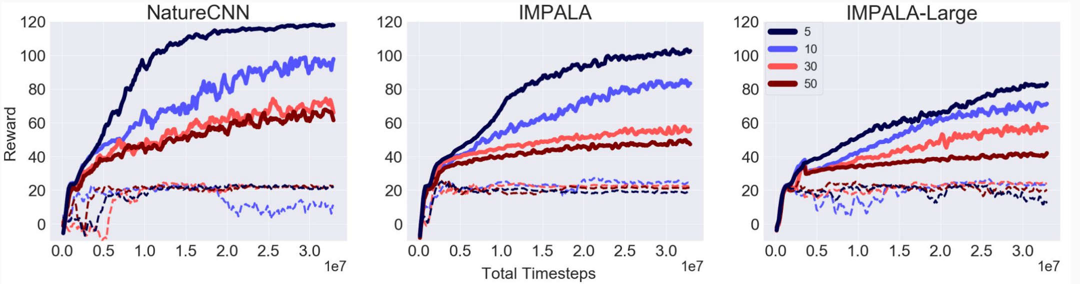
RL Memorization Test

- If we trained NatureCNN (600K params), IMPALA (622K params), and IMPALA-LARGE (823K params) on “the background” g_θ, which policies memorize the most?
- The largest model (i.e. IMPALA-LARGE) should memorize more, right?

RL Memorization Test

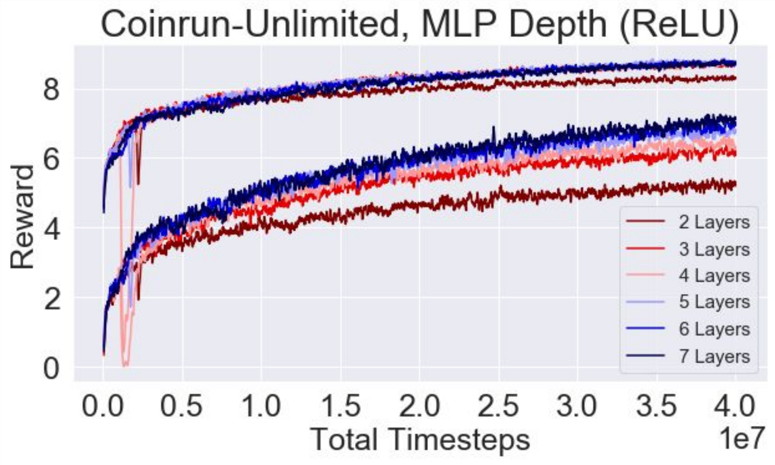
- Nope. IMPALA-LARGE ***memorizes the least!***
- Evidence of **Implicit Regularization** in RL.

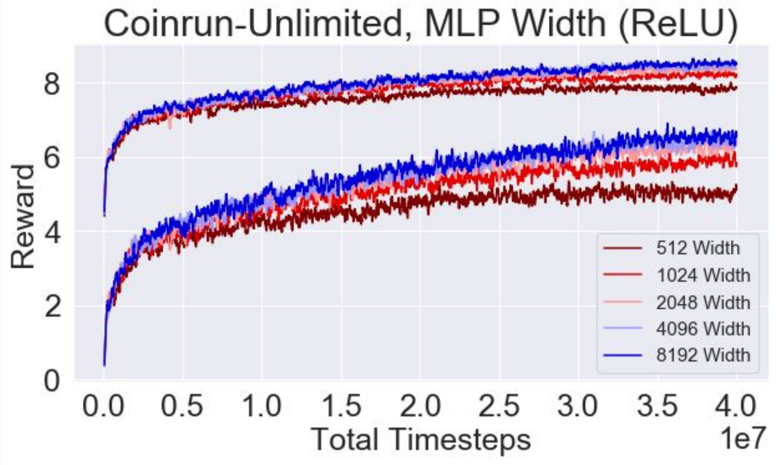




Implicit Regularization in CoinRun

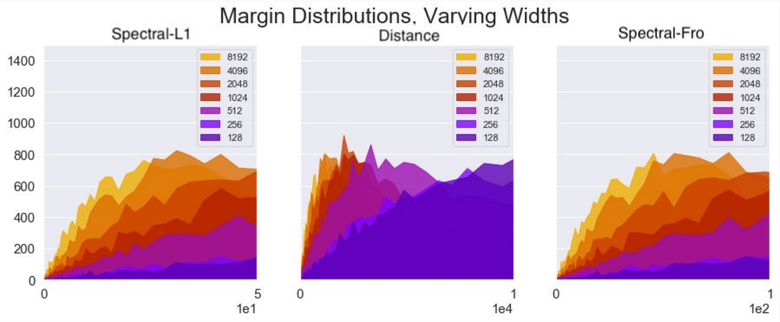
- Does increasing depth/width for MLPs help CoinRun? Yes.

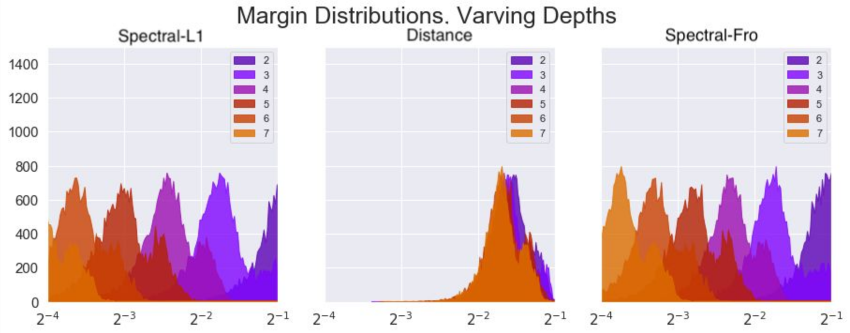




Implicit Regularization in CoinRun

- But are we able to predict generalization gaps at all using classic margin distributions from SL [Bartlett17]?
- Treat on-policy buffer (state, action) pairs as (image, label) pairs in SL.
- Nope. Norm based bounds are too strong.





Conclusions

- Our theoretical understanding of Deep RL generalization is limited.
- SL generalization bounds do not empirically hold at all for RL.
- Overparameterization and Implicit Regularization should be studied more in RL.

Thank you!

Bibliography

- [Gamrian18] - Shani Gamrian, Yoav Goldberg. Transfer Learning for Related Reinforcement Learning Tasks via Image-to-Image Translation. ICML 2019.
- [Zhang18] - Amy Zhang, Yuxin Wu, Joelle Pineau. Natural Environment Benchmarks for Reinforcement Learning. Prerint 2018.
- [Cobbe18] - Quantifying Generalization in Reinforcement Learning. ICML 2019.
- [Zhang17] - Chiyuan Zhang, Oriol Vinyals, Remi Munos, Samy Bengio. A Study on Overfitting in Deep Reinforcement Learning. Preprint 2018.
- [Nichol18] - Alex Nichol, Vicki Pfau, Christopher Hesse, Oleg Klimov, John Schulman. Gotta Learn Fast: A New Benchmark for Generalization in RL. Preprint 2018.
- [Neyshabur17] - Behnam Neyshabur. Implicit Regularization in Deep Learning. PhD Thesis, 2017.
- [Bartlett17] - Peter Bartlett, Dylan Foster, Matus Telgarsky. Spectrally-normalized margin bounds
- for neural networks. NeurIPS, 2017.