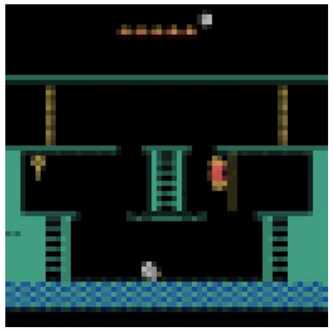
Control-Oriented MBRL with Implicit Differentiation

Evgenii Nikishin Romina Abachi Rishabh Agarwal Pierre-Luc Bacon

Motivation

Standard Model-Based RL agents train dynamics by **reconstructing** next states



Can we learn models that care about **control**?

Differentiate returns w.r.t. model directly!

Implicit Parameterization

Models induce a constraint on values

$$Q(s,a) = B^\theta Q(s,a)$$

What if there's **an implicit function** $\theta \xrightarrow{\varphi} Q^*$

Examples:

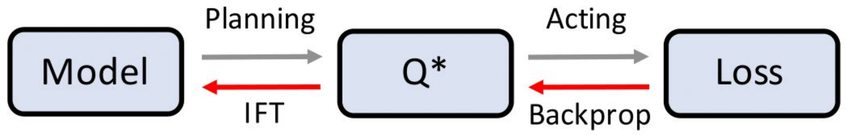
- Value Iteration
- K gradient steps on Bellman error

Derivatives? Implicit Function Theorem!

If $f(\theta, \varphi(\theta)) = \mathbf{0}$ **then**

$$\frac{\partial \varphi(\theta)}{\partial \theta} = -\left(\frac{\partial f(\theta, Q^*)}{\partial Q}\right)^{-1} \cdot \frac{\partial f(\theta, Q^*)}{\partial \theta}$$

Optimal Model Design



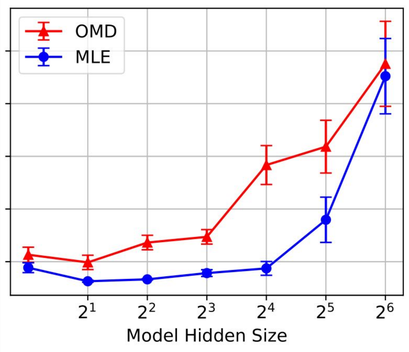
Bi-level optimization:

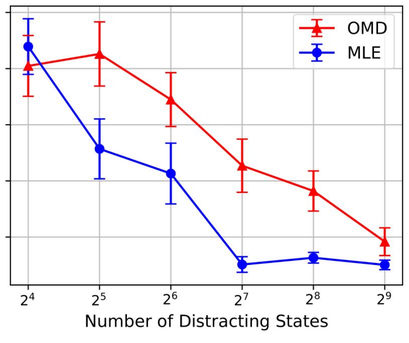
- Inner loop: **get Q* induced by model**
- Outer loop: **optimize real returns w.r.t. model**

Why OMD?

OMD is preferable when:

- True model can't be represented accurately
- Some state components are uninformative





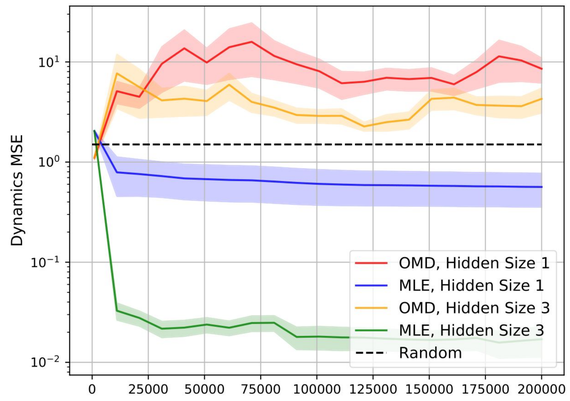
True model is complex → learn what matters!

Implementation Details:

- Approximations to inverse work well
- Approximate Q* works well
- Differs from Dyna only in the model loss
- Easy to implement in JAX

What Does OMD Learn?

- MLE has **high likelihood** but **bad for planning**
- OMD has **low likelihood** but **good for planning**



High likelihood is not needed for a useful model!

Theoretical Analyses

OMD recovers a **Q* equivalent model** θ

$$B^\theta Q^*(s,a) = B^{\text{true}} Q^*(s,a)$$

OMD enjoys a **tighter bound**

$$\max_{s,a} \left| Q^*(s,a) - \hat{Q}_{\text{MLE}}(s,a) \right| \leq \frac{\epsilon_r}{1-\gamma} + \frac{\gamma \epsilon_p r_{\max}}{2(1-\gamma)^2}$$

$$\max_{s,a} \left| Q^*(s,a) - \hat{Q}_{\text{OMD}}(s,a) \right| \leq \frac{\epsilon}{1-\gamma}$$

Takeaways:

1. Learning true models can be hard and not needed
2. OMD: end-to-end control-oriented model learning
3. Search over simpler models that maximize returns