On the Practical Consistency of Meta-Reinforcement Learning

Zheng Xiong, Luisa Zintgraf, Jacob Beck, Risto Vuorio, Shimon Whiteson

UNIVERSITY OF OXFORD

UKRI Engineering and Physical Sciences Research Council

AIMS EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems

Introduction

What is consistency?

A consistent algorithm can find the optimal solution to any test task given an infinite adaptation budget.

Interesting! But why is it important in meta-RL?

RL agents might encounter out-of-distribution (OOD) tasks in real world. Consistency ensures that if needed, the agent can overcome its meta-learned inductive bias for OOD adaptation.

Sounds good. But does it really help *in practice*?

Not sure, as theoretical consistency builds upon strict assumptions that may not hold in practice.

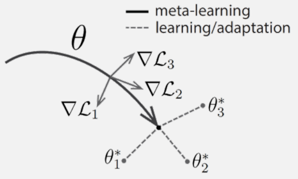
So why not measure it empirically?

That's exactly what we do! This work answers:
Q1: Does theoretical consistency translate into practical benefits when adapting OOD?
A1: Yes, in most cases. But not always!
Q2: If so, can inconsistent algorithms enjoy the same benefits with minor modification?
A2: Yes! We make them practically consistent by continued training on the OOD test tasks.

Methodology

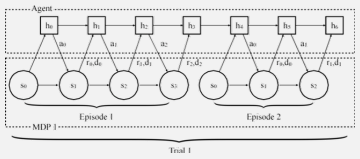
Gradient-based methods

- MAML [1]
- Theoretically consistent 🙂
- Adapt OOD by default

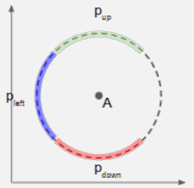
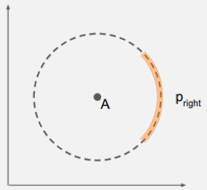
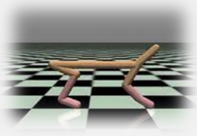


Context-based methods

- *RL*^{2} [2], VariBAD [3]
- Not theoretically consistent 😔
- *What we propose*: continued training (CT) on test tasks enables practical consistency 🙂



Experiments

	Meta-train	Meta-test
2D navigation Navigate to goals in unseen directions		
Mujoco Transfer between different tasks or agent types		

Results

1. On *most* tasks, theoretically consistent methods (MAML) adapt to OOD tasks quite well
2. But theoretical consistency is not always sufficient for practical consistency (E.g. MAML on sparse-navi)
3. Continued training significantly improves practical consistency of context-based methods

Algorithm	Navi	Sparse-navi	Cheetah-vel	Cheetah-vel $\rightarrow$ dir	Cheeetah-dir $\leftrightarrow$ ant-dir
MAML	0.80	0.00	**1.11**	**1.10**	0.73
RL^2-default	-0.49	0.05	-0.06	0.03	-0.01
RL^2-CT	0.52	**0.41**	0.75	0.76	0.39
VariBAD-default	-0.20	0.03	-0.18	0.13	0.04
VariBAD-CT	**0.91**	0.29	0.96	0.99	**0.78**

Table 1. Consistency scores on different tasks. The higher, the better.

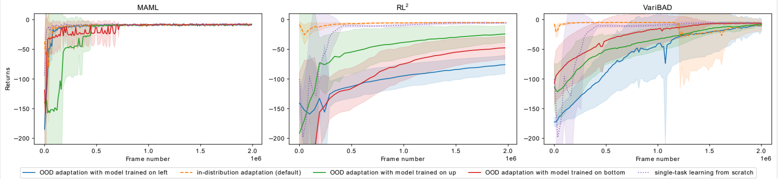


Figure 1. Learning curves of meta-test on navigation to right. The initial scores of *RL*^{2} and VariBAD represent their default performance.

For Further Information

Contact: zheng.xiong@eng.ox.ac.uk

Sorry that I can not present our poster to you in person. But welcome to join our online poster at NeurIPS 2021 workshop on Meta-Learning.

References

[1] Finn, Chelsea, et al. "Model-agnostic meta-learning for fast adaptation of deep networks." International Conference on Machine Learning. PMLR, 2017.
[2] Duan, Yan, et al. "*RL*^{2}: Fast reinforcement learning via slow reinforcement learning." arXiv preprint arXiv:1611.02779 (2016).
[3] Zintgraf, Luisa, et al. "VariBAD: A Very Good Method for Bayes-Adaptive Deep RL via Meta-Learning." International Conference on Learning Representations. 2020.