On the Practical Consistency of Meta-Reinforcement Learning

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

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
• RL² [2], VariBAD [3]
• Not theoretically consistent 😞
• What we propose: continued training (CT) on test tasks enables practical consistency 😊

Experiments

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

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

<table>
<thead>
<tr>
<th>Algorithm</th>
<th>Navi</th>
<th>Sparse-navi</th>
<th>Cheetah-v0</th>
<th>Cheetah-v0 + dir</th>
<th>Cheetah-v0 + ann-dir</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAML</td>
<td>0.00</td>
<td>0.00</td>
<td>1.11</td>
<td>1.10</td>
<td>0.73</td>
</tr>
<tr>
<td>RL² default</td>
<td>-0.49</td>
<td>0.05</td>
<td>-0.06</td>
<td>0.03</td>
<td>-0.01</td>
</tr>
<tr>
<td>RL² CT</td>
<td>0.52</td>
<td>0.44</td>
<td>0.75</td>
<td>0.76</td>
<td>0.39</td>
</tr>
<tr>
<td>VariBAD default</td>
<td>0.20</td>
<td>0.03</td>
<td>-0.18</td>
<td>0.13</td>
<td>0.04</td>
</tr>
<tr>
<td>VariBAD CT</td>
<td>0.91</td>
<td>0.29</td>
<td>0.96</td>
<td>0.99</td>
<td>0.78</td>
</tr>
</tbody>
</table>

Figure 1. Learning curves of meta-test on navigation to right. The initial scores of RL² 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