Note on One-Shot Reinforcement Learning

Alex Dimakis, 5/11/2025

In the last two weeks some papers that have caught my attention: "RL with only one training example" and "Test-Time RL".  
  
In the "One Training example" paper   
the authors find one question and ask the model to solve it again and again. Every time, the model tries 8 times (the Group in GRPO), and a gradient step is performed, to increase the reward which is a very simple verification of the correct answers, repeated thousands of times on the same problem.   
  
The shocking finding is that the model does not overfit to this one question: RL on one example, makes the model better in MATH500 and other benchmarks.   
(If instead you did SFT repeating one training question-solution finetuning, the model would quickly memorize this answer and overfit). But with RL, the model has to solve the problem itself, since it only sees the question, not the answer. Every time it produces different answers, and this seems to prevent overfitting. The other papers are relying on the same phenomenon: you can have a small number of training questions and re-solve them thousands of times. You can do this for the test set (as test-time RL does) and still not overfit. We saw this by doing RL training on half the test set and seeing benefits in the other half.   
  
My thought now is that this shows our RL learning algorithm must be extremely inefficient. If a human is learning by solving a math question, they immediately learn what they can, by solving it once (or twice). No further benefit would come by assigning the same homework problem to students a tenth time. But in RL, we keep asking the model to re-solve the same question thousands of times, and the model slowly gets better. This makes me think that we should be able to have much better RL algorithms since all the information is there, its just not absorbed fast enough.

A graph of a graph of a graph

Description automatically generated with medium confidence

What is the one magical question for 1-example RL?  
Q:"The pressure P exerted by wind on a sail varies jointly as the area A of the sail and the cube of the wind’s velocity V. When the velocity is 8 miles per hour, the pressure on a sail of 2 square feet is 4 pounds. Find the wind velocity when the pressure on 4 square feet of sail is 32 pounds."   
A:(For verification, Answer is 12.8)  
  
its funny that its a physics calculation and gives better benefits in MATH compared to the pure math questions.   
  
What is the relation to "Test-Time RL" (TTRL) paper?  
In TTRL the authors train on the test set. They say we only need the questions in the test set, not the answers because we can compute the answers with majority decoding (which is expected). This is cool but the most important new insight to me is that repeating a small set of questions doesn't overfit.

Does that mean that data doesn't matter? Absolutely not, and the authors of the one-example paper emphasize this:   
"Our work does not imply that large-scale RLVR datasets are unimportant; rather, it emphasizes the value of better data selection and collection for RLVR"  
The new lesson is simply that RL can re-use the same questions without overfitting and that it is probably possible to get much better RL algorithms because humans can learn much faster.

# References

[1] [Reinforcement Learning for Reasoning in Large Language Models with One Training Example, Y. Wang et al, 2025](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/human_like_reasoning/Reinforcement_Learning_for_Reasoning_in_Large_Language_Models_with_One_Training_Example_Wang_2025.pdf)

[2] [TTRL: Test-Time Reinforcement Learning, Y. Zuo et al, 2025](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/human_like_reasoning/TTRL-Test-Time_Reinforcement_Learning_Zuo_2025.pdf)

[3] [Absolute Zero: Reinforced Self-play Reasoning with Zero Data, A. Zhao et al, 2025](https://github.com/dimitarpg13/reinforcement_learning_and_game_theory/blob/main/articles/ReinforcementLearning/human_like_reasoning/Absolute_Zero-Reinforced_Self-play_Reasoning_with_Zero_Data_Zhao_2025.pdf)