Existing Solutions

- **Key idea:** Express the expected return objective as the sum of two expectations inside and outside a memory buffer of sequences $\mathcal{B} = \{ (\mathcal{A}, r_i) \}_{i=1}^{N}$.

$O(\pi) = \sum_{\mathcal{A} \in \mathcal{A}} \pi(\mathcal{A}) R(\mathcal{A})$

- Expectation inside $\mathcal{B}$
- Expectation outside $\mathcal{B}$

- **MAPO** incorporates a memory buffer of promising sequences to compute an unbiased gradient estimate with low variance.

$\nabla_{\pi} O(\pi) = \sum_{\mathcal{A} \in \mathcal{A}} \nabla_{\mathcal{A}} \pi(\mathcal{A}) R(\mathcal{A}) \nabla_{\mathcal{A}} \log \pi(\mathcal{A}) + \sum_{\mathcal{A} \in \mathcal{A}} \nabla_{\mathcal{A}} \pi(\mathcal{A}) R(\mathcal{A}) \nabla_{\mathcal{A}} \log \pi(\mathcal{A})

- Memory weight clipping
  - Force the training to pay attention to the memory by clipping the weight.
  - Trade off bias in the initial stage for faster training.

- Systematic exploration
  - Use a bloom filter to force the exploration to generate new programs.
  - Trade off memory for more efficient exploration.

- Distributed sampling
  - Distribute the cost of computing $\pi_{\mathcal{B}}$ and sampling into the actors.
  - Multiple actors each interacting with a shard of training set and sending samples to a learner to update the model.

- **Experiments**

  - First RL-based state-of-the-art method on WikiTableQuestions.
  - Competitive to state-of-the-art methods on WikiSQL, which use strong supervision (the ground truth programs), while MAPO only uses weak supervision (the final answers).

- **Comparison of MAPO, MML, IML with a simplified example**

- **Which nation won the most silver medal?**

  - Correct program: (argmax rows “Silver”)
    - (hop v1 “Nation”)

  - Spurious programs: (argmax rows “Gold”) (hop v1 “Nation”) (argmax rows “Bronze”) (hop v1 “Nation”)

  - Comparison of MAPO, MML, IML with a simplified example

<table>
<thead>
<tr>
<th>Rank</th>
<th>Nation</th>
<th>Gold</th>
<th>Silver</th>
<th>Bronze</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Japan</td>
<td>1</td>
<td>0</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>2</td>
<td>France</td>
<td>0</td>
<td>1</td>
<td>0</td>
<td>1</td>
</tr>
<tr>
<td>3</td>
<td>Italy</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>4</td>
<td>Germany</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>5</td>
<td>Brazil</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>6</td>
<td>Korea</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>7</td>
<td>Canada</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
<tr>
<td>8</td>
<td>China</td>
<td>0</td>
<td>0</td>
<td>1</td>
<td>1</td>
</tr>
</tbody>
</table>

- Correct program: (argmax rows “Silver”)
  - (hop v1 “Nation”)

- Spurious program: (argmax rows “Gold”) (hop v1 “Nation”)

- MAPO converges slower than maximum likelihood training, but reaches a better solution.

- REINFORCE doesn't make much progress (<10% accuracy).

- Spurious programs: right answer for the wrong reason.

- Which nation won the most silver medal?