Model Checking SS25
Assignment 11
Due: June 17th, 2025, 09:00

For this assignment sheet we will use the shield synthesis tool tempest to test out the effects of shielding on reinforcement learning. We have packed everything needed for the framework into a docker image. The framework consists of multiple parts, namely:

- `MiniGrid`^{SAFE}: A library for reinforcement learning problems,
- `Minigrid2PRISM`: A tool for the automated translation between `MiniGrid`^{SAFE} environments and the `prism` language, and
- `tempest-py`: Python-bindings for `tempest`

Most parts of this framework can be assumed to be working-as-intended and do not need to be modified by you for this exercise. The source files and a dockerfile are available here: https://git.pranger.xyz/sp/MC-ProbMC-PrismFiles/src/branch/main/HW11. Please pull the files from this git repository.
In order to use the framework, there are two options:

1. Build the framework yourself:

   ```
   sudo docker build -t mc:hw11 --build-arg no_threads=6 .
   ```

   This will build the framework using 6 threads and tag the resulting image with `mw:hw11`.

2. We will send out the built image via a seperate email. You can then use the image directly. Note: *This might not work depending on you processor architecture.*

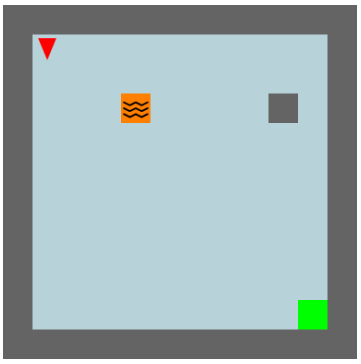
Once you have the image available on your machine, you can start it with the following command:

```
docker run --hostname hw11 -it --rm -v ./workspace:/opt/workspace -v
./Minigrid:/opt/Minigrid mc:hw11
```

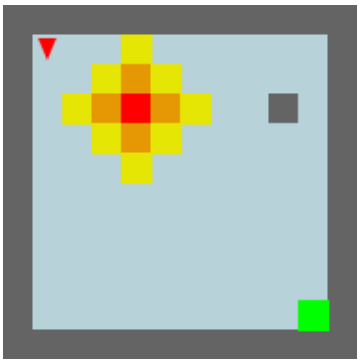
In the running container, you have to activate the virtual environment, install the mounted version of Minigrid^{SAFE} in editable mode and switch to `/opt/workspace`:

```
root@hw11:/opt/Minigrid# source /opt/venv_minigrid/bin/activate
(venv_minigrid) root@hw11:/opt/Minigrid# pip install -e .
# Ensure that the mounted version of Minigrid^SAFE is installed in editable mode
(venv_minigrid) root@hw11:/opt/Minigrid# cd /opt/workspace
```

With the container now running, we can start experimenting. Your primary files to modify are the environment in `./HW11/Minigrid/minigrid/envs/mchw11.py` and `./HW11/workspace/train_minigrid.py`. *You can modify those files outside of the docker container, as their directories are mounted into the container at start-up!*



(a) The environment …



(b) …and a visualization for the shield.

Your task for this assignment sheet is to create an interesting environment for which a shield can successfully prevent an agent from reaching its goal safely, while an unshielded agent fails to do so. The main script for testing this is `./HW11/workspace/train_minigrid.py`. The script handles the translation from your environment to a prism model, as well as the computation of the shield, i.e. the model checking part.
You can run the script for your environment with the following command:

```
python train_minigrid.py --env "MiniGrid-MCHW11Env-v0" --shielding
full
```

Setting the argument for `--shielding` to `none` disables shielding.
While testing, the script produces three main outputs:

- `/opt/workspace/env.png`: An image of your environment, as depicted in Fig. 1a.
- `/opt/workspace/env_and_shield.png`: A visualization of your shield, as depicted in Fig. 1b. This visualization follows the coloring as seen in the lecture: Yellow indicates that some actions are disallowed, orange means that only fallback actions are allowed, and red means that the property has been violated.
- `/opt/workspace/training_results/**/progress.csv`: A csv containing the logged training results.

The assignment sheet is subdivided into four tasks:

Task 1: **[20 Points]** Modify the class `MCHW11Env` by adding more lava pools (an optionally walls) to create an interesting environment. You are also allowed to move the goal position. Hand in your `env.png`

Task 2: **[10 Points]** Run the training script with shielding disabled and hand in the resulting `progress.csv`. This should show that the agent is accumulating safety violations.

Task 3: **[30 Points]** Set parameters of the threshold $F <= n$, with $n < 30$, and the `shield_value`, such that the shield completely blocks the agent from successfully reaching the goal. Please hand in at least one `progress.csv` alongside a description of the parameters and the resulting `env_and_shield.png`.

Task 4: **[40 Points]** Set parameters of the threshold $F <= n$, with $n < 30$, and the `shield_value`, such that the shield successfuly prevents the agent from running into the lava. Please hand in at least one `progress.csv` alongside a description of the parameters and the resulting `env_and_shield.png`.

Have fun!