Supplemental Information:
Robustness assessment of energy policies: Reinforcement learning approach to support myopic energy transition

Xavier Rixhon[1,2,*], Hervé Jeanmart[1], and Francesco Contino[1]

[1]Institute of Mechanics, Materials and Civil Engineering (iMMC), Université catholique de Louvain (UCLouvain), Place du Levant, 2, 1348 Louvain-la-Neuve, Belgium
[2]Lead contact
[*]Correspondence: xavier.rixhon@uclouvain.be

The Supplemental Information begins with

Supplementary Note 1 Reinforcement Learning fundamentals and algorithm

Reinforcement Learning (RL) is a subfield of machine learning focused on training an agent to make sequential decisions by interacting with an environment to achieve specific goals (see Figure 1). Unlike supervised learning, where data is labelled, and unsupervised learning, where patterns are inferred from unlabelled data, reinforcement learning deals with learning from interaction, typically through trial and error. This way, RL is considered as active learning[1]. Starting from an initial state, the agent takes an action that impacts its environment. The latter feeds back the agent with a reward and the new state. This goes on until the end of the episode. When the episode is done, the agent starts again from an initial state, takes an action and so on.

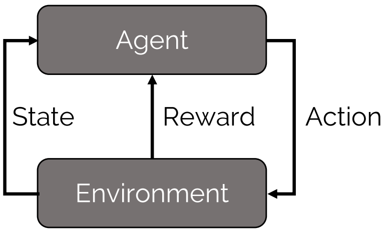

Figure 1: General concept of Reinforcement Learning (RL) as the interactions between the agent and its environment. The agent takes some action that has an impact on the environment which feeds back the agent with a reward and the new state. The objective of the agent is to optimise its policy, i.e., the mapping between the state it is at and the action to take, by maximising its cumulative reward.

The agent learns to optimise its policy by maximising the cumulative reward over time. This policy refers to the strategy or mapping from states to actions that the agent employs to make decisions. Essentially, it defines the behaviour of the agent in the environment. The ultimate goal of the agent is often to find an optimal policy, which maximises the expected cumulative reward over time. All these concepts and interactions between the agent and its environment are formalised as a Markov Decision Process (MDP)[2], represented by the tuple $< s, a, T, r, \pi, \gamma >$. The Markov property of such a decision process states that a decision is made only based on this tuple and not on the history/path that has led to it. In this tuple, $s \in S$ is the state defined in a certain state space, S, that represents the observable parts of the environment that the agent uses to make decisions; $a \in A$ is the action among the action space, A; T is the probability of transitioning from one state s to another state s' given a specific action, a: $T(s, a, s')$: $\Pr(s'|s, a)$; r is the reward received by the agent when taking the action a from state s, $r(s, a)$; π is the policy telling the action to take depending on the current state and; γ is the discount factor that controls the importance of future rewards versus immediate rewards. During the learning/optimisation process, the agent acts according to the exploitation-exploration trade-off. In the exploitation, the action a is directly given by the mapping provided by the current policy π, depending on the state s. In the exploration, the action is randomly picked within the action space. For further information, the interested reader is invited to refer to the work of Sutton and Barto[2] or the course given by David Silver[3] available online.

Before jumping to the choice of the learning algorithm, it is worth noting that we opted for the combination of RL with Deep Neural Network (DNN), called Deep Reinforcement Learning (DRL). Among others, one of the main drawbacks of traditional RL algorithms, i.e., without the use of Neural Network (NN), is that it suffers from the “curse of dimensionality” when facing problems with continuous action and state spaces (see Chapter **??**). By approximating the state-action function with its parameters (i.e., weights and biases), DNN can address this difficulty.

Given the assumed absence of knowledge of the agent about the dynamics of the environment, i.e., its transition or reward functions, we needed a so-called “model-free” learning algorithm. In Reinforcement Learning, the “model” stands for the dynamics “action-state-reward” between the environment and the agent. In practice, in a model-free approach, the agent estimates the optimal policy directly from experience and without estimating the dynamics of the environment. However, model-free methods suffer from two major drawbacks: their sample inefficiency and their sensitivity concerning their hyper-parameters (e.g., learning rates, exploration constants)[4]. The former leads to a too-expensive computational burden while the second requires meticulous settings to get good results. To overcome these two challenges, we needed to choose between an “on-policy” or “off-policy” algorithm. In a nutshell, in on-policy learning, the agent learns the value function or policy based on the data it generates by following its current policy whereas, in off-policy, the agent can learn from data collected by any policy, not just the one it is currently following, which provides greater flexibility and potential for reusing data stored in the so-called replay buffer. This makes off-policy algorithms more data efficient and ensures better exploration by reusing past experiences or even following random exploration[4].

To optimise the mapping between the observations and the actions, the policy $\pi\left(\boldsymbol{a}_n|\boldsymbol{o}_n\right)$, an objective function, $\boldsymbol{J}(\pi)$, is built on the cumulative rewards collected during each episode. Finally, a back-propagation process updates the weights and biases of the NN during the learning of the agent. Among the wide variety of RL algorithms applied in energy systems[5], this work opted for Soft Actor-Critic (SAC)[4] to train and update the NN. Like other actor-critic-based algorithms, SAC works with two NN in parallel: the actor learning the control policy and the critic judging the actor (see Figure 2).

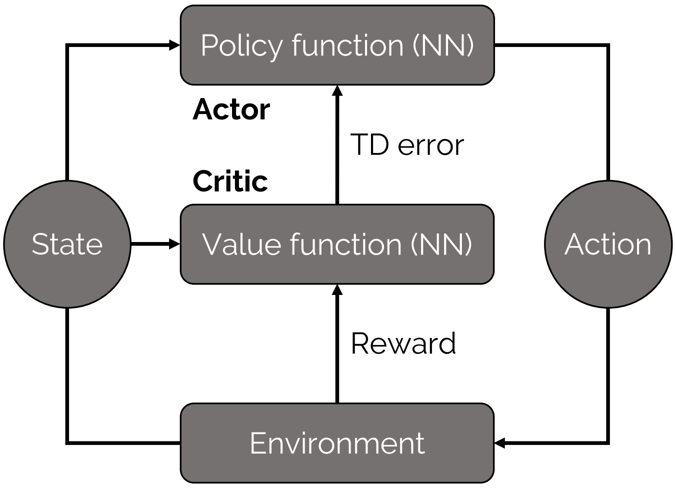

Figure 2: General concept of actor-critic-based algorithms. The two NN are trained against each other for the actor to improve the control policy and for the critic to provide a better judgement of the actor’s action via the temporal-difference (TD) error. Graph adapted from[1].

SAC is a model-free and off-policy actor-critic deep RL algorithm based on the entropy-augmented objective function (see Equation (1)). The word “augmented” here is in opposition to the conventional RL objective function that is only based on the cumulative reward, i.e., the first term of Equation (1). In the RL context, entropy, also called “Shannon entropy”, stands for the randomness or stochasticity of the policy.

$$\boldsymbol{J}(\pi) = \mathop{\mathbb{E}}_{\pi}\left[\sum_{n=0}^{N_{ep}} \gamma^n r_n\left(\boldsymbol{o}_n, \boldsymbol{a}_n\right) - \zeta \log\left(\pi\left(\boldsymbol{a}_n | \boldsymbol{o}_n\right)\right)\right], \tag{1}$$

where γ is the discount factor and ζ the temperature parameter. γ determines how much importance we want to give to future rewards within an episode. ζ balances the trade-off between the exploitation of proven actions via the return maximisation, i.e., $\sum_{n=0}^{N_{ep}} \gamma^n r_n\left(\boldsymbol{o}_n, \boldsymbol{a}_n\right)$, and the exploration through the entropy term, i.e., $\log\left(\pi\left(\boldsymbol{a}_n | \boldsymbol{o}_n\right)\right)$. This way, SAC ensures sample efficiency while improving exploration[6] and robustness[7]. In their work, Haarnoja et al.[6] showed a lower sensitivity of SAC to hyper-parameters. These make SAC a state-of-the-art algorithm and one of the most efficient model-free deep RL methods nowadays[6]. In this thesis, we used the open-source SAC package developed by STABLE-BASELINES3[8] where the policy NN is a fully connected multilayer perceptron (MLP) built with TENSORFLOW[9]. For further information on RL and the SAC algorithm, the interested reader is invited to refer to the works of Sutton and Barto[2] and Haarnoja et al.[4], respectively.

Supplementary Note 2 Comparison with perfect foresight under uncertainties

This section compares these results under myopic conditions supported by the RL-agent with the perfect foresight under uncertainties that is considered as a reference.

RL-based myopic optimisation provides CO_2 emissions pathways different from the perfect foresight approach to respect the same CO_2 budget (see left side of Figure 3). However, driven first by this CO_2 budget, the agent often reaches much lower cumulative emissions when succeeding the transition (see Figure **??**). This comes from the agent's actions that limit the emissions and/or the consumption of fossil resources at the early stages. Thanks to the bigger emission reduction at these early stages, the RL-based optimisation can benefit from a "CO_2 buffer" at the end of the transition. This buffer is compensated by the end of the transition where 50% of the myopic transitions reach 2050 with 10 or more remaining Mt$_{\mathrm{CO_2,eq}}$ compared to 4 for the perfect foresight approach. These remaining emissions by 2050 come from the consumption in industrial boilers of waste and coal accounting for 3.5% and 2.4% of the primary mix on average by 2050. Finally, the long-term vision of the perfect foresight approach results in a smoother reduction of emissions to end up with less emissions by 2050.

The comparison between the failures and the successes demonstrates the added value brought by myopic pathway optimisation. In the near term (2025-2030), levels of emission are similar between perfect foresight and myopic cases that have failed. This shows that limited foresight encourages to strongly act at the early stages. On top of this, following the initial steps of CO_2 emissions pathways resulting from the PF approach would likely ($\sim$80%) lead to failure of the transition.

Looking at the total transition cost, the combination of the agent's actions and favourable economic conditions (see Section **??**) make the myopic transitions cheaper, on average, than the PF cases (see right side of Figure 3). This is also because the perfect foresight approach always finds a solution even in the worst conditions such as the high cost of purchasing resources and high End-Use Demand (EUD). This explains the wider variability of the PF results too. However, with the same sample of uncertain parameters, given the assumed full knowledge over the whole time horizon, PF naturally results in a cheaper transition than its myopic equivalent.

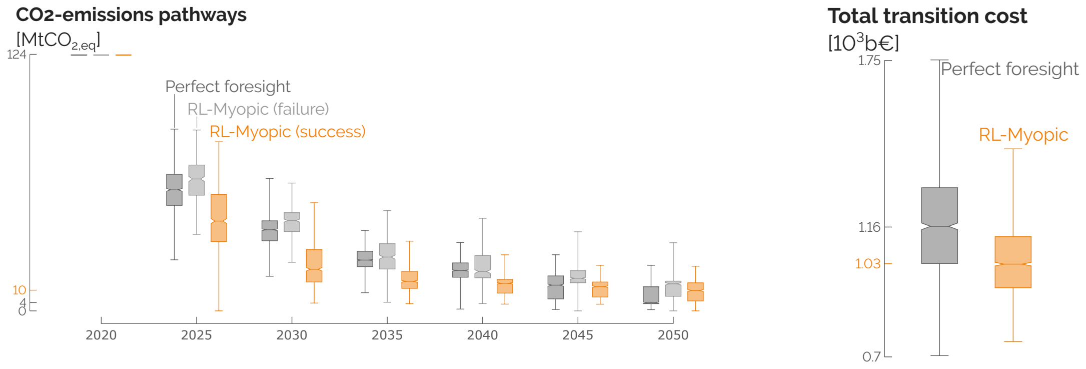

Figure 3: Comparison of CO_2 emissions pathways (left) and total transition cost (right) from the perfect foresight optimisation under uncertainties and the RL-based myopic optimisation. Myopic transitions succeed with a more drastic reduction of emissions in the short term and, on average, more favourable economic conditions.

The analysis of the cumulative costs shows that the Operational Expenditure (OPEX) is the main difference between myopic and perfect foresight transitions (see Figure 4). Supported by the agent's actions, successful myopic transitions have a lower OPEX than the perfect foresight ones.

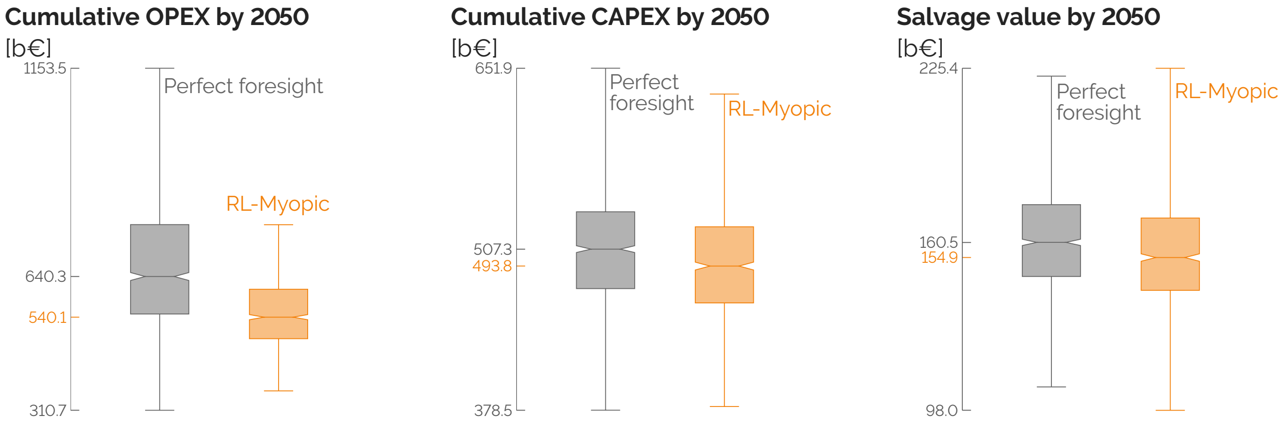

Figure 4: Comparison of cumulative OPEX (left), CAPEX (centre) and salvage value (right) in 2050 from the perfect foresight optimisation under uncertainties and the RL-based myopic optimisation.

The cost of purchasing the energy carriers represents about 70% of the total cumulative OPEX. The assessment of the primary energy mix by 2050 highlights that the difference in OPEX between the perfect foresight and the myopic pathways comes from the import of electrofuels, and especially of e-ammonia (see Figure 5). In the majority of the cases, e-ammonia is more than two times more imported in the myopic transitions. Being cheaper than e-methane (see Chapter **??**), e-ammonia brings flexibility in the production of electricity via Combined Cycle Gas Turbine (CCGT) (see Chapter **??**). Besides the slightly favourable economic conditions (see Table **??**), the myopic optimisations opt to invest massively into importing renewable molecules because of the limited knowledge of the future, and, among others, the availability of Small Modular Reactor (SMR). This explains why 50% of the successful transitions reached cumulative emissions below 900 Mt$_{\mathrm{CO_2},\mathrm{eq}}$ (see Figure **??**).

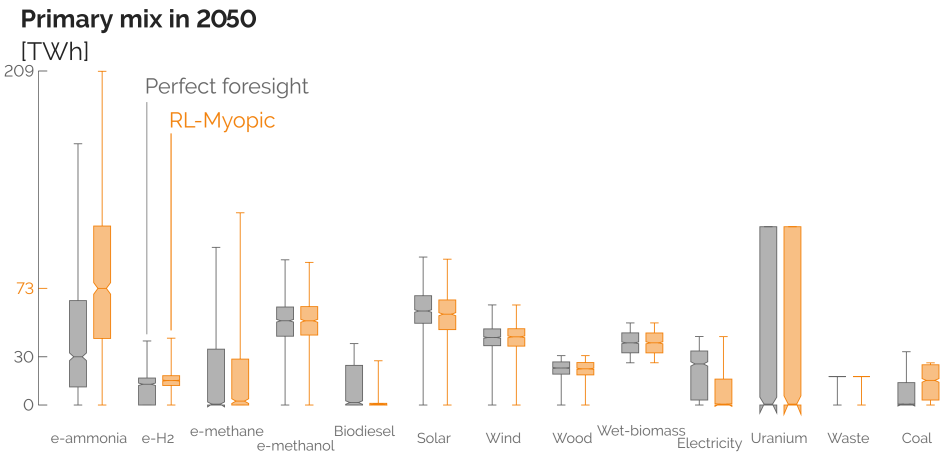

Figure 5: Comparison of the primary energy mix in 2050 from the perfect foresight optimisation under uncertainties and the RL-based myopic optimisation. The biggest difference is about e-ammonia to supply CCGT.

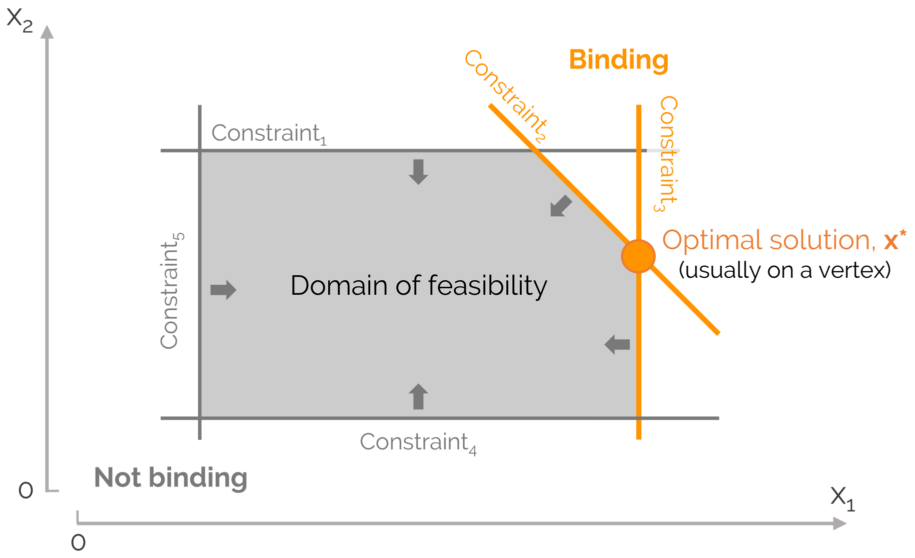

Figure 6: Binding versus non-binding constraints. In LP where the feasibility domain is non-empty and bounded, the constraints defined a convex feasibility domain in the space of variables (here, x_1 and x_2). The optimal solution usually locates on a vertex of this domain, i.e., the intersection of several constraints (here, constraints 2 and 3) limiting the solution. These constraints are considered binding, i.e., having a limiting impact on the optimal solution.

Supplementary Note 3 To bind or not to bind, that is the question

To identify the actions that have an actual impact on the environment, we can check if they are binding or not. In a Linear Programming (LP) problem, constraints represent hyperplanes in the domain of variables. In a two-dimension space, these are straight lines (see Figure 6). When the problem is bounded and feasible, these lines are the edges of a convex polygon: the domain of feasibility. The optimal solution, $\mathbf{x}^*$, is the combination of variables leading to the optimal value of the objective function. Besides being within the domain of feasibility, it is proven that this optimal solution, when unique[1], locates on a vertex of the domain[10]. The constraints intersecting at this vertex are considered binding, actually limiting the objective function to be more optimal. In other words, binding constraints, when tightened, aggravate the objective value function. If these are inequality constraints, as represented in Figure 6, it means that the left and right sides of the equations are equal.

References

1. Cao, D., Hu, W., Zhao, J., Zhang, G., Zhang, B., Liu, Z., Chen, Z., and Blaabjerg, F. (2020). Reinforcement learning and its applications in modern power and energy systems: A review. Journal of modern power systems and clean energy *8*, 1029–1042.

2. Sutton, R.S., and Barto, A.G. (2018). Reinforcement learning: An introduction. MIT press.

[1] There are cases where the objective function has the same optimal value along an entire edge. In this case, there is an infinity of solutions and the problem is indeterminate.

3. David Silver (2016). RL Course by David Silver. `https://www.youtube.com/watch?v=2pWv7GOvuf0&list=PLzuuYNsE1EZAXYR4FJ75jcJseBmo4KQ9-.`.

4. Haarnoja, T., Zhou, A., Abbeel, P., and Levine, S. (2018). Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In International conference on machine learning. PMLR pp. 1861–1870.

5. Perera, A., and Kamalaruban, P. (2021). Applications of reinforcement learning in energy systems. Renewable and Sustainable Energy Reviews *137*, 110618.

6. Haarnoja, T., Tang, H., Abbeel, P., and Levine, S. (2017). Reinforcement learning with deep energy-based policies. In International conference on machine learning. PMLR pp. 1352–1361.

7. Ziebart, B.D. (2010). Modeling purposeful adaptive behavior with the principle of maximum causal entropy. Carnegie Mellon University.

8. Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. (2021). Stable-baselines3: Reliable reinforcement learning implementations. The Journal of Machine Learning Research *22*, 12348–12355.

9. Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G.S., Davis, A., Dean, J., Devin, M. et al. (2016). Tensorflow: Large-scale machine learning on heterogeneous distributed systems. arXiv preprint arXiv:1603.04467.

10. Bertsimas, D., and Tsitsiklis, J.N. (1997). Introduction to linear optimization vol. 6. Athena Scientific Belmont, MA.