Guided A* Search for Scheduling Power Generation Under Uncertainty

Patrick de Mars [1] **Aidan O'Sullivan** [1]

Abstract

Increasing renewables penetration motivates the development of new approaches to operating power systems under uncertainty. We apply a novel approach combining self-play reinforcement learning (RL) and traditional planning to solve the unit commitment problem, an essential power systems scheduling task. Applied to problems with stochastic demand and wind generation, our results show significant cost reductions and improvements to security of supply as compared with an industry-standard mixed-integer linear programming benchmark. Applying a carbon price of \$50/tCO$_2$ achieves carbon emissions reductions of up to 10%. Our results demonstrate scalability to larger problems than tackled in existing literature, and indicate the potential for RL to contribute to decarbonising power systems.

1. Introduction and Related Work

The power sector is the single largest contributor to global CO$_2$ emissions (Ritchie & Roser, 2020) and requires rapid decarbonisation to achieve climate goals. One of the fundamental problems in power systems operation is determining the on/off (commitment) schedules of generation to meet demand, the unit commitment (UC) problem. The UC problem is traditionally solved by mixed-integer linear programming (MILP), with a reserve constraint enforced to manage uncertainties. However, in high renewables systems, such deterministic methods are outperformed by stochastic formulations (Ruiz et al., 2009), but these have seen limited real-world application due to high computational costs and other practical challenges (Bertsimas et al., 2012).

In this paper, we tackle the day-ahead UC problem with a combination of self-play reinforcement learning (RL) and traditional planning methods. A policy is trained by self-play RL in a power system environment, and applied at test time to reduce the branching factor of a search tree which is solved by A* search (Russell & Norvig, 2009). We use 'guided A* search' to solve UC problems with stochastic demand and wind generation, using forecasts based on GB power system data. We consider power systems of 10, 20 and 30 generators, each with and without a carbon price to understand: (1) cost-competitiveness with an MILP benchmark; (2) scalability to larger power systems; (3) the impact of carbon pricing on carbon emissions. Compared with MILP, guided A* search achieves lower operating costs and loss of load probability, with no loss of performance with increasing problem size. Implementing a carbon price shifts generation away from coal-fired power stations towards less carbon intensive generation, reducing total carbon emissions by up to 10%. Our research contributes a novel and scalable approach applying RL to the UC problem, and shows significant potential to learn complex operational strategies and outperform MILP approaches.

Existing research in this area (Jasmin & TP, 2009; Dalal & Mannor, 2015; Jasmin et al., 2016; Navin & Sharma, 2019; Li et al., 2019) has focused on small numbers of generators, in part due to the combinatorial action space that limits the application of existing RL methods 'out-of-the-box'. In the most similar research (Dalal & Mannor, 2015), tree search methods are applied to a system of 12 generators, which, to the best of our knowledge, is the largest prior study in this area. However, the problem considered is deterministic and does not consider generalisability to unseen problems. In subsequent related research, a larger power system is considered but the UC problem is simplified to a single commitment decision per day (Dalal et al., 2016). To the best of our knowledge, our research is unique in testing on unseen profiles and investigating carbon pricing.

In the next section, the UC problem is formulated as a Markov Decision Process (MDP), suitable for RL methods. We present our methodology of guided A* search in Section 3, which we apply in experiments described in Section 4. We discuss our results and conclude the paper in Section 5.

2. Unit Commitment as an MDP

To apply RL methods, we formulate the UC problem as an episodic MDP (Sutton & Barto, 2018), with episodes consisting of 48 decision periods reflecting half-hour market settlement periods. At each timestep, the agent receives

[1] UCL Energy Institute, London, UK. Correspondence to: Patrick de Mars <patrick.demars.14@ucl.ac.uk>.

Tackling Climate Change with Machine Learning Workshop at ICML 2021.

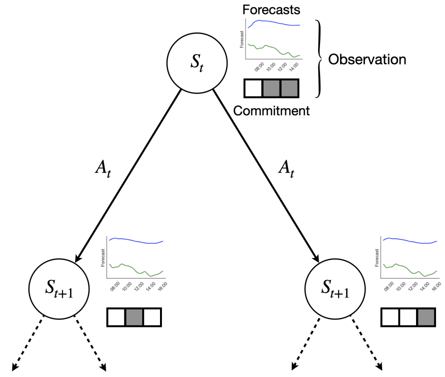

Figure 1. Example search tree representing the UC problem for a system of 3 generators. Nodes on the tree represent observations, comprised of demand and wind forecasts and generator up/down times. Actions are combinatorial decisions to commit or decommit generators at the next timestep. The step cost of traversing an edge is the expected operating cost.

an observation consisting of the following components: (1) current generator up/down times $\boldsymbol{u}_t$; (2) demand forecast $\boldsymbol{d}_t$; (3) wind forecast $\boldsymbol{w}_t$. Demand and wind forecast errors x_t and y_t are included in the state s_t but unobserved by the agent. The errors are sampled from error distributions X_t and Y_t, which are modelled using with auto-regressive moving average (ARMA) processes (as used elsewhere in power systems literature (Soder, 2004)). An action $a_t \in \{0, 1\}^N$ is chosen by the agent, determining the on/off status for each of N generators (subject to generator constraints) at the next timestep. The environment processes a_t by evaluating the transition function $F(s_{t+1}, s_t, a_t)$, updating the generator up/down times, sampling forecast errors and rolling the forecasts forward one timestep. The realisations of net demand (demand minus wind generation) are used to calculate the 'economic dispatch' determining the real-valued power outputs p_i for each generator i, such that $\sum_i p_i$ equals the net demand, if possible. Using the economic dispatch, CO_2 emissions are calculated along with fuel costs, carbon emissions costs, startup costs and lost load costs (i.e. a penalty when there is not enough capacity to meet net demand). The negative sum of all costs is the reward r_t. The agent aims to maximise the discounted return from the initial state $\mathbb{E}[\sum_{t=0}^{T-1} \gamma^t R_{t+1}]$ with discount factor γ.

3. Guided Tree Search Algorithms

Since the environment dynamics are largely known and can be modelled, planning methods can be applied to solve the MDP. In contrast with model-free RL, using the model to predict the outcome of actions is very valuable given the importance of safe operation; lost load events (which may lead to blackouts) carry extreme penalties in the reward function and are catastrophic in the real world.

We formulate the UC MDP as a search tree, where nodes represent states and edges represent actions (Figure 1). Although the transition function is stochastic, such that there is a one-to-many mapping from $(s_t, a_t) \rightarrow s_{t+1}$, observations omit the stochastic components of the state (forecast errors x_t and y_t) meaning there is a one-to-one mapping in the observation space. The search tree can therefore be constructed in the observation space. Under this formulation, the cost of traversing an edge (negative reward) is stochastic, depending on realisations of random variables X_t and Y_t. The edge costs are set using a Monte Carlo method, sampling the transition many times and computing the mean cost. Due to the combinatorial nature of the action space, ordinary planning methods such as A* search (Russell & Norvig, 2009) are infeasible, with exponential time complexity in the number of generators. We apply a novel technique called guided expansion, which exploits an expansion policy $\pi(a|s)$ to choose a subset of actions to add to the search tree. This enables planning on a reduced search tree. We train the expansion policy using self-play RL, in an open-source environment developed for this research[1].

3.1. A* Search

We use the informed search method A* search (Russell & Norvig, 2009) to find the lowest cost path through the search tree, applying two adjustments to the original algorithm. First, in order to prevent an exponential explosion in the computation time with respect to the depth, we apply real-time A* (Korf, 1990), repeatedly using A* to solve a reduced sub-tree limited to a lookahead horizon H. For each decision period, the truncated sub-problem is solved, and the first child in the solution path is chosen as the root for the next sub-problem. This algorithm has linear time complexity in the number of decision periods T. Second, rather than fix H to be constant, we apply iterative deepening (Korf, 1985) and incrementally increase H beginning at $H = 1$, while a computational time budget b is available. This makes our algorithm *anytime*, meaning the search can be terminated at any point and return a solution for the sub-problem. In practice, this is an appealing characteristic, as electricity market constraints mean that there is a limited timeframe (typically of the order of minutes) in which the UC problem must be solved. Applying iterative deepening maximises the search depth within the computational budget.

3.2. Guided Expansion

Even with the real-time and iterative deepening adjustments, applying A* search to even moderately large power systems is infeasible, since the number of branches can be up to 2^N for N generators. We use a novel method that we call

[1] https://github.com/pwdemars/rl4uc

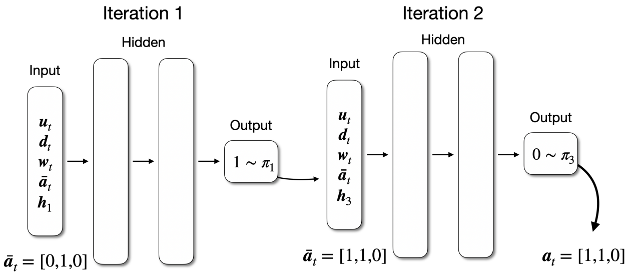

Figure 2. Sequential feed-forward neural network architecture used to parameterise the expansion policy. Each generator commitment is classified sequentially with the current action sequence $\bar{a}_t$ used to estimate the following commitment. In the example, the second generator is constrained to remain on, so at the first iteration, $\bar{a}_t = [0, 1, 0]$.

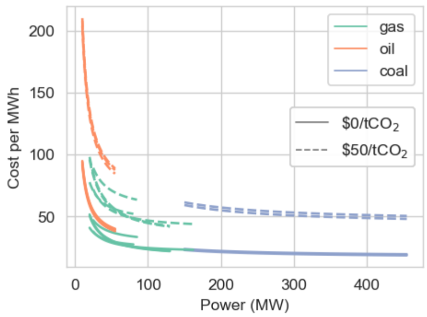

Figure 3. Quadratic cost curves for the 10 generators specified in (Kazarlis et al., 1996), with and without carbon pricing.

'guided expansion' to reduce the branching factor of the search tree. When adding nodes to the tree, an 'expansion policy' $\pi(a|s)$ (which we train using model-free RL) selects a subset of actions, by adding only those actions which satisfy $\pi(a|s) \geq \rho$, where $0 \leq \rho < 1$ is fixed a branching threshold. The maximum number of nodes M that can be added to the tree is therefore limited to $M \leq \frac{1}{\rho}$.

We train the expansion policy with self-play RL using proximal policy optimisation (PPO) (Schulman et al., 2017) on a set of training episodes. Fully enumerating the actions at the output layer of the policy is not feasible due to the size of the action space. Parametrising the multi-dimensional action space with N output nodes is also not appropriate due to the strong dependency of each generator's action propensity on the other generators. Instead, the policy is parametrised as a binary classifier which sequentially predicts each value in the sequence $\boldsymbol{a} = [a_1, a_2, ..., a_N]$ representing an action, where $a_i \in \{0, 1\}$ are sub-actions giving the commitment for generator i (Figure 2). The output of the classifier at each iteration is passed as an input into the next forward-pass through the network, thus maintaining the history of generator commitments already decided. In addition, the input vector includes a one-hot encoding indicating the a_i being classified on each forward pass as well as the observation. This parametrisation succeeds in preserving the interdependencies between generators while remaining tractable for larger power systems.

3.3. Priority List Heuristic for Unit Commitment

A* search exploits a domain-specific heuristic $h(n)$ to estimate the least cost from a node n to a goal node (cost-to-go). We designed a heuristic based on a simple priority list (PL) algorithm, which commits generators in order of their minimum operating cost. To reduce computation time, this heuristic only considers the initial minimum up/down time constraints. The PL heuristic provides a 'best-case' (*admissible*) estimate for the operating costs, allowing for sub-trees to be effectively pruned and enabling deeper search within the time budget.

4. Experiments

We conducted two experiments: (1) comparing the performance of guided A* with a MILP benchmark with no carbon price; (2) investigating the impact of a carbon price. For each experiment, we investigated power systems of 10, 20 and 30 generators to evaluate scalability, using generator specifications from (Kazarlis et al., 1996). The data define quadratic fuel cost curves, minimum and maximum power limits, minimum up/down time constraints and startup costs for each generator. The penalty for lost load was set to \$10,000/MWh, based on estimates in (Schröder & Kuckshinrichs, 2015). The demand and wind forecasts used for training and testing episodes were based on real data from the GB power system (2016–2019), retrieved from (BMRS, 2021), with 20 days withheld for testing. In training, the wind penetration (wind generation as proportion of demand) was 17%, with a maximum daily penetration of 58%.

The expansion policies for each power system were trained asynchronously with PPO over 8 CPU workers. These policies were then used in guided A* search to solve 20 unseen test episodes, with time budgets b between 2–60 seconds per decision period. We set the branching threshold $\rho = 0.05$ for all problems, limiting the branching factor to 20. Each UC solution was evaluated using Monte Carlo simulations, calculating the economic dispatch and associated costs, loss of load probability and other metrics under 1000 realisations of demand and wind forecast errors sampled from the ARMA processes. This allows for comparison of relative performance *on average*, considering multiple realisations of uncertainties for each episode.

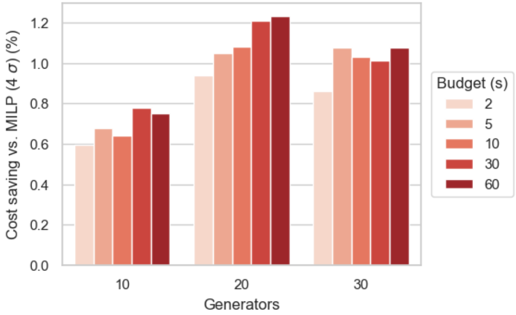

Figure 4. Operating cost saving using guided A* compared with MILP. Savings generally improve with time budget which allows for deeper search.

4.1. Guided A* Search vs. MILP

The first experiment considered power systems with no carbon price. We compared guided A* with a typical MILP benchmark, employing a reserve constraint. We used Power Grid Lib open-source software to solve the UC MILP formulation defined in (Knueven et al., 2020)[2]. The reserve constraint was set to 4σ, where σ is the long-run standard deviation of the net forecast error $(X_t - Y_t)$. This is a typical industry technique for determining reserve constraints (Holttinen et al., 2008).

Guided A* achieved operating cost savings of up to 0.8%, 1.2% and 1.1% for the 10, 20 and 30 generators, respectively (Figure 4). In general, performance improved with increasing time budget, enabling deeper search. The loss of load probability (LOLP) for guided A* search was roughy 50% lower than for MILP, representing more secure operation of the power system. Guided A* search employs more extreme actions (switching multiple generator commitments at once) than MILP, demonstrating complex operational strategies. For instance, in the 20 generator problem guided A* search switches up to 9 generators at once, compared with up to 5 for MILP.

4.2. Impact of Carbon Pricing

In the second experiment, generators were assigned gas, oil or coal fuel types with emissions factors of 54, 73, 95 kgCO$_2$/MMBTU respectively and a carbon price of \$50/tCO$_2$ was applied to fuel use. The cost curves with and without the carbon price applied are shown in Figure 3. The larger capacity generators are coal-fired in our experiments, having the highest emissions factors, and also have the longest minimum up/down time constraints. By contrast, the oil-fired generators have the lowest capacity and shortest minimum up/down times.

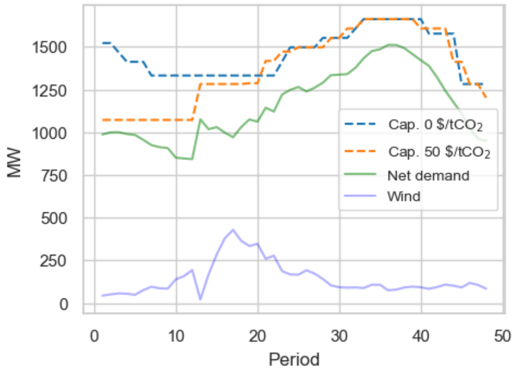

Figure 5. Comparison of capacity committed (cap.) using guided A* search for 10 generator test profile 2016-01-12, with and without a carbon price applied. With a carbon price applied, guided A* search operates smaller reserve margins to improve efficiency of the online generators.

Table 1. Comparison of operational characteristics for A* with and without a carbon price of \$50/tCO$_2$. Coal, gas and oil power stations up-times are shown (% of all periods spent online).

# Gens	\$/tCO$_2$	LOLP (%)	ktCO$_2$	Coal (%)	Gas (%)	Oil (%)	Startups
10	0	0.12	264.03	99.64	41.88	6.28	141
10	50	0.12	245.89	91.30	61.37	13.19	114
20	0	0.11	527.56	99.09	40.74	8.21	235
20	50	0.09	476.62	86.38	66.24	5.38	164
30	0	0.16	780.43	99.10	40.89	5.69	346
30	50	0.17	724.81	88.59	67.86	12.67	215

Applying the carbon price caused significant operational changes and carbon emissions reductions of 6.9, 9.7 and 7.1% for 10, 20 and 30 generator problems respectively (Table 1). The emissions reduction is primarily caused by a shift from the highest carbon intensity baseload units towards gas-fired power stations with significantly lower carbon intensity. The medium carbon intensity oil-fired peaking units still see limited use to manage fluctuations in demand, although total startups decreased. Figure 5 compares the capacity committed by guided A* search with and without the carbon price applied for one test episode, showing tighter reserve margins when the carbon price is applied. LOLP did not increase substantially with the carbon price, despite the tighter margins.

5. Conclusion

The increasing share of renewables required to decarbonise the electricity generation mix demands new UC solution methods that more rigorously account for uncertainties. We have shown that combining RL with planning methods is a viable and scalable methodology for solving the UC problem, achieving cheaper and more secure operation than the MILP benchmark. Sequential parametrisation of the policy

[2] https://github.com/power-grid-lib/pglib-uc

was important to achieving tractability and overcoming the curse of dimensionality in the action space that has limited previous research to systems of up to 12 generators. By contrast, we did not observe a decrease in performance as we increased the problem size to 30 generators.

Shaping the reward function with a carbon price resulted in carbon emissions savings of between 7–10% and changes to operational behaviour, such as fewer startups. Coal generation was displaced by gas, reminiscent of the recent phase-out of coal in the GB power system. Due to the growing size and ubiquity of power systems, such changes in the usage of existing generation assets can yield significant reductions in global CO_2 emissions.

Acknowledgements

We thank Prof. Zoltan Nagy, who provided valuable feedback and mentorship during the writing of this paper. We acknowledge the use of UCL's Myriad High Performance Computing cluster for this research. This research was supported by an Engineering and Physical Sciences Research Council research studentship (grant number: EP/R512400/1).

References

Bertsimas, D., Litvinov, E., Sun, X. A., Zhao, J., and Zheng, T. Adaptive robust optimization for the security constrained unit commitment problem. *IEEE Transactions on Power Systems*, 28(1):52–63, 2012.

BMRS. Balancing Mechanism Reporting Service. https://www.bmreports.com, 2021.

Dalal, G. and Mannor, S. Reinforcement learning for the unit commitment problem. In *2015 IEEE Eindhoven PowerTech*, pp. 1–6. IEEE, 2015.

Dalal, G., Gilboa, E., and Mannor, S. Hierarchical decision making in electricity grid management. In *International Conference on Machine Learning*, pp. 2197–2206. PMLR, 2016.

Holttinen, H., Milligan, M., Kirby, B., Acker, T., Neimane, V., and Molinski, T. Using standard deviation as a measure of increased operational reserve requirement for wind power. *Wind Engineering*, 32(4):355–377, 2008.

Jasmin, E. and TP, I. A. Reinforcement learning solution for unit commitment problem through pursuit method. In *2009 International Conference on Advances in Computing, Control, and Telecommunication Technologies*, pp. 324–327. IEEE, 2009.

Jasmin, E., Ahamed, T. I., and Remani, T. A function approximation approach to reinforcement learning for solving unit commitment problem with photo voltaic sources. In *2016 IEEE International Conference on Power Electronics, Drives and Energy Systems (PEDES)*, pp. 1–6. IEEE, 2016.

Kazarlis, S. A., Bakirtzis, A., and Petridis, V. A genetic algorithm solution to the unit commitment problem. *IEEE Transactions on Power Systems*, 11(1):83–92, 1996.

Knueven, B., Ostrowski, J., and Watson, J.-P. On mixed-integer programming formulations for the unit commitment problem. *INFORMS Journal on Computing*, 32(4): 857–876, 2020.

Korf, R. E. Depth-first iterative-deepening: An optimal admissible tree search. *Artificial Intelligence*, 27(1):97–109, 1985.

Korf, R. E. Real-time heuristic search. *Artificial Intelligence*, 42(2-3):189–211, 1990.

Li, F., Qin, J., and Zheng, W. X. Distributed q-learning-based online optimization algorithm for unit commitment and dispatch in smart grid. *IEEE transactions on cybernetics*, 50(9):4146–4156, 2019.

Navin, N. K. and Sharma, R. A fuzzy reinforcement learning approach to thermal unit commitment problem. *Neural Computing and Applications*, 31(3):737–750, 2019.

Ritchie, H. and Roser, M. CO2 and Greenhouse Gas Emissions. *Our World in Data*, 2020. https://ourworldindata.org/co2-and-other-greenhouse-gas-emissions.

Ruiz, P. A., Philbrick, C. R., Zak, E., Cheung, K. W., and Sauer, P. W. Uncertainty management in the unit commitment problem. *IEEE Transactions on Power Systems*, 24 (2):642–651, 2009.

Russell, S. and Norvig, P. *Artificial Intelligence: A Modern Approach*. Prentice Hall Press, USA, 3rd edition, 2009. ISBN 0136042597.

Schröder, T. and Kuckshinrichs, W. Value of lost load: An efficient economic indicator for power supply security? a literature review. *Frontiers in Energy Research*, 3:55, 2015.

Schulman, J., Wolski, F., Dhariwal, P., Radford, A., and Klimov, O. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

Soder, L. Simulation of wind speed forecast errors for operation planning of multiarea power systems. In *2004 International Conference on Probabilistic Methods Applied to Power Systems*, pp. 723–728. IEEE, 2004.

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