PowRL: A Reinforcement Learning Framework for Robust Management of Power Networks

Anandsingh Chauhan[1], **Mayank Baranwal**[1,2], **Ansuma Basumatary**[3]

[1]Tata Consultancy Services Research, Mumbai
[2]Indian Institute of Technology, Bombay
[3]SalesKen, Bengaluru
anandsingh.chauhan@tcs.com, baranwal.mayank@tcs.com, ansuma.bty25@gmail.com

Abstract

Power grids, across the world, play an important societal and economical role by providing uninterrupted, reliable and transient-free power to several industries, businesses and household consumers. With the advent of renewable power resources and EVs resulting into uncertain generation and highly dynamic load demands, it has become ever so important to ensure robust operation of power networks through suitable management of transient stability issues and localize the events of blackouts. In the light of ever increasing stress on the modern grid infrastructure and the grid operators, this paper presents a reinforcement learning (RL) framework, PowRL, to mitigate the effects of unexpected network events, as well as reliably maintain electricity everywhere on the network at all times. The PowRL leverages a novel heuristic for overload management, along with the RL-guided decision making on optimal topology selection to ensure that the grid is operated safely and reliably (with no overloads). PowRL is benchmarked on a variety of competition datasets hosted by the L2RPN (Learning to Run a Power Network). Even with its reduced action space, PowRL tops the leaderboard in the L2RPN NeurIPS 2020 challenge (Robustness track) at an aggregate level, while also being the top performing agent in the L2RPN WCCI 2020 challenge. Moreover, detailed analysis depicts state-of-the-art performances by the PowRL agent in some of the test scenarios.

Introduction

The infrastructure that defines the modern electricity grid is largely based on centralized power generation units, such as the fossil-fuel-fired power plants and nuclear power plants (Wulf 2000). Through out its evolution, the primary focus of grid development has largely been around safety, reliability, and resiliency to uncertain events. The ever increasing penetration of centralized renewable generation units, such as the hydroelectric dams, as well as the distributed energy resources (DERs), such as the solar or wind farms, has resulted in significant variability in power generation. Moreover, the customers' evolving energy usage patterns through charging of electric vehicles (EVs), batteries, and accessibility to modern electrical appliances have furthered the uncertainty and variability in the electric grid, not to mention the

occasional adversarial cyberattacks in order to disrupt reliable power delivery by causing blackouts.

Modernization of the electricity grid through innovative hardware and software solutions involving peer-to-peer networks of power electric converters for enabling coordinated response to adjust power generation and consumption can partially alleviate some of the challenges (Baranwal et al. 2018; Arwa and Folly 2020). However, overhauling the entire grid infrastructure, unfortunately, operates at its own time-scale. This work addresses the challenge of leveraging AI-driven solutions for extracting the maximum possible resiliency out of the current grid infrastructure in order to ensure smooth power delivery and provide ancillary services to the grid operator. The challenge lies not only in dealing with the uncertainty of power demand and generation, or the uncertain events, such as electrical faults or adversarial attacks on the grid, but also with the huge (combinatorially large) action space even in a moderately-sized grid. From the perspective of the grid operator, devising a real-time strategy for the robust management of power networks is beyond human cognition. In most such scenarios, the grid operator relies on his/her own experience or at best, some of the potential heuristics whose scope is limited to mitigating only a certain types of uncertainties.

Motivating example Fig. 1 represents a typical scenario in a power network, where one of the transmission lines (shown in red) is overloaded beyond its thermal capacity. If left unaccounted, the overflow may result in permanent damage of the transmission line resulting in significant delays in getting the network back to its nominal state. Interestingly, an optimal and real-time control strategy is to identify substations where a bus-split action (as shown in the rightmost grid) results in immediate mitigation of line overflow through appropriate changes in the underlying network topology. Finding the most-suited topology control action among the list of all possible topologies is not straightforward, especially if the power network is equipped with many substations with each substation consisting of multiple control elements.

Why RL? The penetration of DERs, EVs, and complex connectivities between different substations is increasingly rendering the traditional control strategies used by the electrical engineers inadequate. On the other hand, recent advancements in machine learning (ML) may offer solutions to such complex problems which are otherwise beyond hu-

man cognition. A key feature of most power systems is that the operating point, characterized by the net power generation and load demand, is highly dynamic. Consequently, managing a power network reliably entails real-time sequential decision making, for which reinforcement learning (RL) is most suited. Given the potential solutions that an RL-based framework offers, the "Learning to run a power network" (L2RPN) challenge was recently conceptualized to model the sequential decision-making environments of real-time power network operations. The challenge introduced a realistically-sized network environment (based on the IEEE-118 network), which is a reduced-order approximation of the American Electric Power system (in the U.S. Midwest). L2RPN is aimed at testing the feasibility of an RL-based approach towards developing a smart operation recommender system for grid operators.

Statement of Contributions

In this work, we introduce an RL-based recommender system, titled *PowRL*, for robust management of power networks subjected to adversarial attacks on transmission lines (uncertain events), as well as temporal variations in demand and supply. PowRL is built upon the *Grid2Op* (Donnot 2020a) framework that facilitates users to plug in the underlying simulator of a model of their choice. Below we summarize the primary contributions of our work.

1. **Grid management with reduced action-space** The total number of feasible topology actions in the reduced IEEE-118 system are in the order of 70k. Working with such a large action space is beyond the scope of any practical RL-agent. Following the work in (Zhou et al. 2021; Yoon et al. 2021), we identify important substations and a significantly reduced action set (comprising of 240 unique actions) through extensive simulations in order to facilitate faster learning by agents.

2. **Combining heuristics with deep-learning** Traditional control of electrical grids is often based on a supervisory rule-based strategy. We augment a novel rule-based scheme with the RL agent during both training and inference phases in order to develop a coordinated control response to uncertain events (see Fig. 3).

3. **SOTA performance on benchmark datasets** PowRL is benchmarked on several L2RPN competition datasets, and is shown to significantly outperform the classical heuristics, as well as other RL agents.

4. **Post-hoc analysis and insights** We also analyze the sequences of control actions adopted by the PowRL agent across various scenarios, and observe that unlike other RL agents, the action sequences of PowRL are more diversified suggesting improved learning behavior.

Preliminaries

A power network comprises of several *substations* connected with each other through transmission *lines*. Each substation is equipped with electrical elements, such as *generators* and/or *loads*. For instance, the network in Fig. 1 consists of 4 substations (nodes), 5 transmission lines (edges), 2 generators ($P1$ and $P2$), and 2 loads ($P3$ and $P4$). The

generators produce power, only to be consumed by the network loads and through transmission losses. The power lines are subjected to maximum current (power) carrying capacities, and an excess flow through power lines for a sizable duration results in permanent damage and eventual disconnection from the network. A substation acts as a router in the network and determines where to transmit power. Additionally, a substation is also equipped with multiple conductors, known as *buses*. The last subfigure in Fig. 1 represents a scenario where the network at the fourth substation is routed through 2 separate buses.

The Grid2Op Environment

The L2RPN challenge is built upon an open-source simulator, *Grid2Op*, for power grid operation. Grid2Op offers flexibility to work with realistic scenarios, execute contingency events, and test control algorithms subjected to several physical and operational constraints. The challenge scenario consists of industry standard synthetic IEEE-118 network with 36 substations, 59 lines and 22 generators (see Fig. 2). The remainder of the IEEE-118 network is represented as a time-varying load. Grid2Op considers 2 buses per substation, also known as the double busbar system. Grid2Op environment is also equipped with realistic production and consumption scenarios, as well as adversarial attacks on the network manifested as forced line failures. The agents can take remedial actions subjected to following constraints:
(a) Deterministic events, such as maintenance, and adversarial events, such as meteorological conditions, can disconnect lines for substantial duration
(b) Each power line has a maximum flow capacity, and gets disconnected automatically if overloaded for too long
(c) Disconnected lines cannot be immediately reconnected
(d) Agents cannot act too frequently on the same line, generator or substation in order to avoid asset degradation
The episode terminates when the total load demand is not met or during incidents of wide-area blackout

States, Actions and Rewards

The problem of controlling power networks can be cast as a Markov Decision Process (MDP). An MDP is denoted by the tuple $\langle \mathcal{S}, \mathcal{A}, \mathcal{P}_\mathcal{A}, r \rangle$, where $\mathcal{S}$ and $\mathcal{A}$ represent the finite set of states and actions, respectively. For each $s, s' \in \mathcal{S}$, the probability of transition from $s \rightarrow s'$ under the effect of an action $a \in \mathcal{A}$, is denoted by $p_a(s, s') \in \mathcal{P}_\mathcal{A}$. Finally, the step-reward associated with each state-action pair (s, a) is depicted by $r(s, a)$. Below, we summarize the set of all possible states, actions and reward:
States Agents can access the entire state of the power network at each time-step, including the demand forecast at the next step, load flow and status of power lines, voltages at each busbar, production at each generator, and various operational constraints
Actions Grid2Op is equipped with two kinds of actions: (a) combinatorial (or discrete), (b) continuous. Discrete actions are related to inexpensive topological actions, such as line disconnection/reconnection, or actions at the busbar. Additionally, the generators can be redispatched through predefined continuous actions using costly production changes.
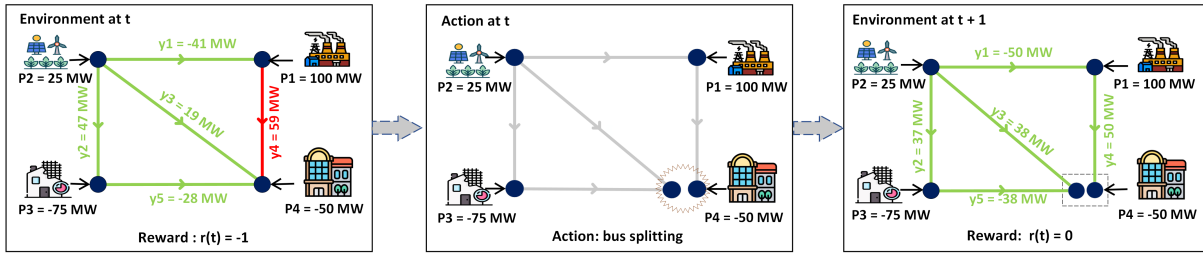
Figure 1: Reliable operation through bus-splitting: instead of disconnecting the overloaded line, a simpler and effective solution is to alter the underlying network topology.
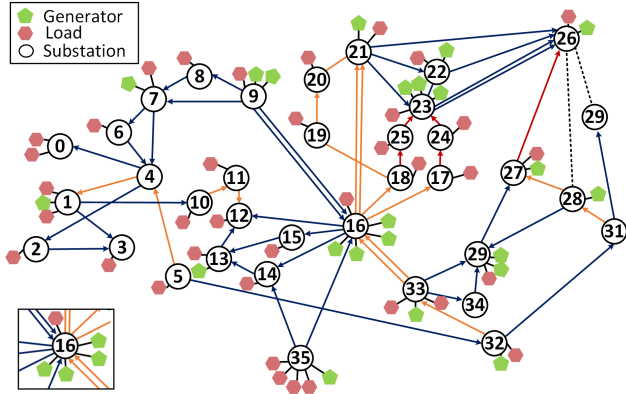

Figure 2: Synthetic IEEE-118 network consisting of 36 substations, 59 lines, 22 generators.

For the reduced IEEE-118 system, there are nearly 70k discrete actions and 40 continuous actions.

Rewards The L2RPN challenge is equipped with a competition specific reward, however, Grid2Op also facilitates inclusion custom reward functions. More details on the exact reward structure are included in (Marot et al. 2021).

Challenges in Controlling Power Network

Combinatorially many actions The primary challenge with executing remedial topological reconfiguration actions is that the number of discrete actions at a substation scales exponentially with the number of elements connected to the substation. For instance, the substation 16 in Fig. 2 comprises of 17 elements (12 lines, 4 generators, 1 load). The number of topologies at this substation for a double busbar system is 65,505. The Proposition below enumerates all valid topologies for a double busbar substation.

Proposition 1. *The number of valid topological reconfigurations for a double busbar substation comprising of N_{line} lines, N_g generators, and N_{load} loads is $2^{N_{tot}-1} - 2^{N_g+N_{load}} + 1$, where $N_{tot} := N_{line} + N_g + N_{load}$.*

Proof: See Technical Appendix for details.

Uncertainty with renewables Renewable energy sources, such as solar or wind, depend heavily on the weather conditions. Thus, while the current topology may be adept at handling the net load demand, a significant change in power generation due to change in weather conditions may force the operator to take immediate remedial action in order to avoid transmission loss failures or eventual blackouts.

Adversarial attacks The L2RPN challenge operates with a heuristically designed opponent to mimic the $N-1$ security criterion in power networks (Zhao et al. 2018); it acts in an environment parallel to the RL agent and affects the power grid through forced contingencies, by disconnecting some of the targeted tensed lines at random times. The lines remain disconnected for maintenance and the agent can only reconnect them once the maintenance is over.

Related Works

Among the most standard practices for managing contingency events is the '$N-1$'-criterion (Ren, Dobson, and Carreras 2008), which requires that the network should continue to operate in the safe state even if one of the elements (productions, lines, transformers, etc.) is disconnected. The criterion is usually enforced through solving a complex optimization problem that encode several physical constraints, as well as the fundamental laws of electricity (Wang and McCalley 2013; Dehghanian et al. 2015; Alhazmi et al. 2019). Due to the limitations with the traditional approaches, there is a shift towards adopting AI guided approaches as potential solution methodologies for robust control of power networks. One of the foundational works along this direction (Donnot et al. 2017) introduces a deep-learning based approach for remedial actions comprising of line disconnection/reconnection in a fixed-topology network. As such, their approach does not explore combinatorial topological actions. In the subsequent works, the authors proposed an expert system for topological remedial actions (Marot et al. 2018) for the IEEE-14 system, which was further improved using a simple RL-based framework (Marot et al. 2020) through introduction of the Grid2Op simulator and the first power grid management competition in 2019.

The winner of the first challenge (Lan et al. 2020), used pre-training to generate a good optimal control policy, followed by guided exploration in the large action space to select a top-few legal control actions at each time step through extensive simulations. They designed a dueling deep Q-network (DDQN) (Wang et al. 2016) with prioritized replay buffer (Schaul et al. 2016), only to act in contingency situations. Authors in (Yoon et al. 2021) adopted a graph neural network (GNN) based actor-critic method to find the goal topology of the network given the current state. Their solution, though, was ranked first in the L2RPN WCCI challenge, did not perform nearly as well when applied to more complex scenarios as part of the L2RPN NeurIPS challenge.

The last L2RPN NeurIPS 2020 challenge saw new RL-based solutions that were significant improvements over the previously reported approaches (Marot et al. 2021). The third-place winner, **lujixiang** (Zhihong et al. 2020), used a DDQN approach for topological actions, very similar to (Lan et al. 2020), however, with a reduced action space of 885 actions. The authors also included a base agent to handle remedial actions involving line disconnection/reconnection and generator redispatch. The second-place winner, **binbinchen** (Hu, Chen, and Tang 2021), used a novel 'Teacher-Tutor-Junior-Senior-Student' framework for warm-start their proximal policy optimization (PPO) (Schulman et al. 2017) model comprising of 208 unique remedial actions identified through extensive simulations. The top performing entry, **rl_agent**, from Baidu (Zhou et al. 2021) relied on their PARL framework and Baidu's computational resources to train a very deep policy using evolutionary black-box optimization (Lee et al. 2020).

Our Approach: PowRL

Obtaining Reduced Action Space

Even with AI-guided solutions, working with very large action space (comprising of nearly 70k actions) is impractical. More importantly substations, such as the ones indexed by 9, 16, 23 and 26 (Fig. 2) consist of relatively larger number of control elements, and understandably offer greater flexibility in terms of control of power networks. Through extensive simulations, we identify a much smaller of 240 actions that appeared to have most impact on network control, i.e., $|\mathcal{A}| = 240$. The action set consists of only topological control. Costly generation production actions are excluded in our approach, as the inexpensive topological changes alone responded well to several contingency events.

Heuristic-Guided Topological Actions

Other than building an RL agent to select a top-performing (electrically) valid action from the set $\mathcal{A}$, we also introduced some baseline heuristics to support the RL agent. These heuristics include: (a) disengage RL agent during 'acceptable' grid operation, (b) manually disconnect a line during sustained periods of overflow in order to avoid permanent damage, as well as allowing PowRL to reconnect the line back soon after the cooldown period ends, (c) any network reconfigurations are restored to original state as soon as the contingency ends, (d) reconnect the line back soon after the scheduled maintenance period is over, (e) do not disconnect lines that result in network bifurcation.

Proximal Policy Optimization (PPO)-Guided Reinforcement Learning Framework

In order to learn to perform optimal sequential decision making, we train an on-policy PPO agent with a prioritized replay buffer defined over the set of network states. The state of the agent $S_t \in \mathcal{S}$ at time-step t consists of: (a) the time-step information t including the month, date, hour, minute and day of week information. (b) generator features F_{gen} comprising of active (P_{gen}) and reactive (Q_{gen}) generation, and the voltage magnitude V_{gen}. (c) load features

F_{load} comprising of active (P_{load}) and reactive (Q_{load}) consumption, and the voltage magnitude V_{load}. (d) line features F_{line} which include active ($P_{\text{or}}, P_{\text{ex}}$) and reactive ($Q_{\text{or}}, Q_{\text{ex}}$) power flows, as well as the bus voltages ($V_{\text{or}}, V_{\text{ex}}$) and current flow ($a_{\text{or}}, a_{\text{ex}}$), at both ends of the line. The other important line features are the power flow capacity ρ a ratio to capture the amount of overflow in a line, timestep overflow t_{of} to depict the duration of time steps since a line has overflowed, and line status I_{l} to indicate if the line is connected or disconnected in the network. (e) Additionally, the state space also includes the topological vector status $I_{\text{topo\_vect}}$ to indicate on which bus each object of the power grid is connected in the substation. (f) The grid operator may have to wait for some time steps before a line or a substation can be acted upon, denoted by the line cooldown time step t_{l} and substation cooldown time step t_{s}. (g) Power lines can go under some planned maintenance; the information of the time of next planned maintenance t_{nm} and its duration t_{d} is also included in the input features, and as an agent can't operate on lines under maintenance. In summary, the state of the PPO agent is given by:

$$S_t := [t, F_{\text{gen}}, F_{\text{load}}, F_{\text{line}}, I_{\text{l}}, \rho, I_{\text{topo\_vect}}, t_{\text{of}}, t_{\text{l}}, t_{\text{s}}, t_{\text{nm}}, t_{\text{d}}].$$

The step reward for training the PPO agent is designed to incur additional penalty when the maximum overflow ratio ρ_{max} is beyond the safety threshold of 0.95, i.e.,

$$r = \begin{cases} 2 - \rho_{\text{max}}, & \text{if } \rho_{\text{max}} < 0.95, \\ 2 - 2\rho_{\text{max}}, & \text{else.} \end{cases}$$

Additionally, we incentivize the agent through an episodic reward of 500 upon surviving the entire episode. Premature termination of the episode due to any illegal action, or grid blackout is penalized through an episodic reward of -300.

Optimal Action Selection

Controlling a power network using random exploratory actions may result in power grid collapse/blackout within a few steps. On the other hand, the `do-nothing` action that does not make any modifications to the existing grid topology or to the generator production, is no good either. In fact, the expected survival rate with the do-nothing action on the L2RPN NeurIPS 2020 challenge set is only $\sim 24\%$. Recall that under nominal operating conditions, it is preferable to continue to operate with the starting topology. This makes it challenging to decide when to sample actions from the RL agent so that the deviation from the initial grid topological configuration is the least for better survival. We achieve this by combining heuristics with the action recommendation by the RL agent. Fig. 3 explains the structured optimal action selection of the PowRL framework. Given the current state S_t of the power system and overflow status of all power lines, the agent first simulates (not implements) the `do-nothing` action and obtains the simulated $\rho_{\text{do-nothing}}$. If $\rho_{\text{do-nothing}}$ is within the safety threshold, then the agent implements either the `do-nothing` action or the `recovery` action in case the current topology is different from the initial topology. In an event of overflow with $\rho_{\text{do-nothing}}$ exceeding the safety threshold, the PowRL framework implements the action recommended by the RL agent in conjunction with
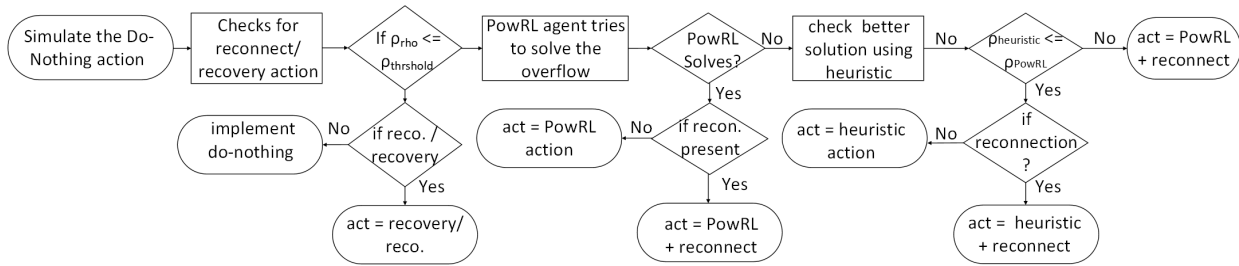
Figure 3: Schematic of PowRL: PowRL combines RL-agent with a threshold based heuristic scheme

Method	NeurIPS (Online)	NeurIPS (Offline)	WCCI	# actions ($|\mathcal{A}|$)	Run-time (s)
Do_Nothing	0	0.34	50.09	NA	0.54
expert_heuristic	22.53	51.44	50.1	NA	9.76
lujixiang	45	53.96	NA	885	824.58
binbinchen	52.42	3.96	51.1	208 + 1255	503.88
rl_agent	61.05	**61.06**	96.4	232 + 500 + redispatch	358.82
PowRL (Ours)	**61.48**	59.69	96.4	**206 + 34**	**353.36**

Table 1: Performance on L2RPN Challenge Datasets

any possible line reconnection action. The performance of the PowRL agent depends on the choice of safety threshold $\rho_{\text{threshold}}$. In scenarios where the network is under severe attack from the adversaries, the PowRL framework needs to perform aggressive topological reconfiguration actions by frequently invoking the RL agent. This can be achieved by setting a smaller value of the safety threshold.

Experiments

Datasets The dataset used for training PowRL is part of the small dataset included with the L2RPN NeurIPS 2020 challenge (Robustness track) starting kit (Donnot 2020b). It consists of power grid data worth 48 years spaced out at an interval of 5 minutes amounting to a total training data of 4,644,864 steps. The dataset is based on the reduced IEEE-118 system (see Fig. 2) with 22 generators, 36 substations, 37 loads, and 59 power lines. some of the loads in this dataset represent interconnection with another grid. This dataset uses the Grid2Op framework to facilitate sequential decision-making process.

The L2RPN NeurIPS 2020 (Robustness track) uses two different datasets - (a) an *offline* dataset included with the starting kit for participants to evaluate the performance of their trained models, (b) an *online* dataset that is hidden from the participants. The participants can directly upload their trained models on the competition Codalab to get them evaluated and ranked. Both these datasets contain 24 episodic weekly scenarios resolved at a 5-minute interval. The hidden dataset was carefully picked to offer the different levels of toughness that might not have been observed during the training phase, even though both the datasets were generated from a similar statistical distribution. The previous version of this competition, the L2RPN WCCI 2020 challenge, differs only in terms of an adversary opponent. While the online submissions to the WCCI challenge are no longer

available, an offline evaluation dataset is included with the WCCI starting kit (Donnot 2020c). This third test dataset consists of 10 different 3-day-long (864 steps) episodic scenarios, picked in order to offer different levels of difficulty. In this work, we benchmark the PowRL framework on all three publicly available datasets.

Experimental setup PowRL runs on Grid2Op, which uses the Gym interface to interact with an agent. This Grid2Op platform emulates the sequential decision-making in the power system, where each episode is divided into a list of states each corresponding to a time step of 5 minutes. Each of these states is described by the power flow at any given time step, which in turn is described by the amount of electricity supplied or consumed by the generators and loads, respectively. This data on power flow in each state is provided in the form of power network architecture, consumption, and generation by the RTE. We have also used the lightsim2grid (Donnot 2020d) backend to the Grid2Op platform in order to accelerate the computation. The model was trained on a large shared facility comprising of 3 Tesla P100 GPUs, with a maximum job duration of 120 hours.

Neural network architecture and training For the PowRL framework, an RL agent based on clipped PPO was trained on the L2RPN NeurIPS robustness track data. As the name suggests, the clipped PPO architecture (Schulman et al. 2017) clips the objective function (ratio between the new and the old policy scaled by the advantages), to remove incentives if the new policy gets too far from the older policy. The clipped PPO architecture consists of actor and critic networks; the actor produces the logits/ probability distribution of all possible actions, and the critic evaluates the goodness of the action suggested by the actor. For the actor network, we used four fully connected layers (with ReLU activation) of dimensions $[1000, 1000, 1000, 208]$. The output layer is a linear layer of a dimension equal to the number of discrete actions (2 out of 208 actions were deemed illegal since they
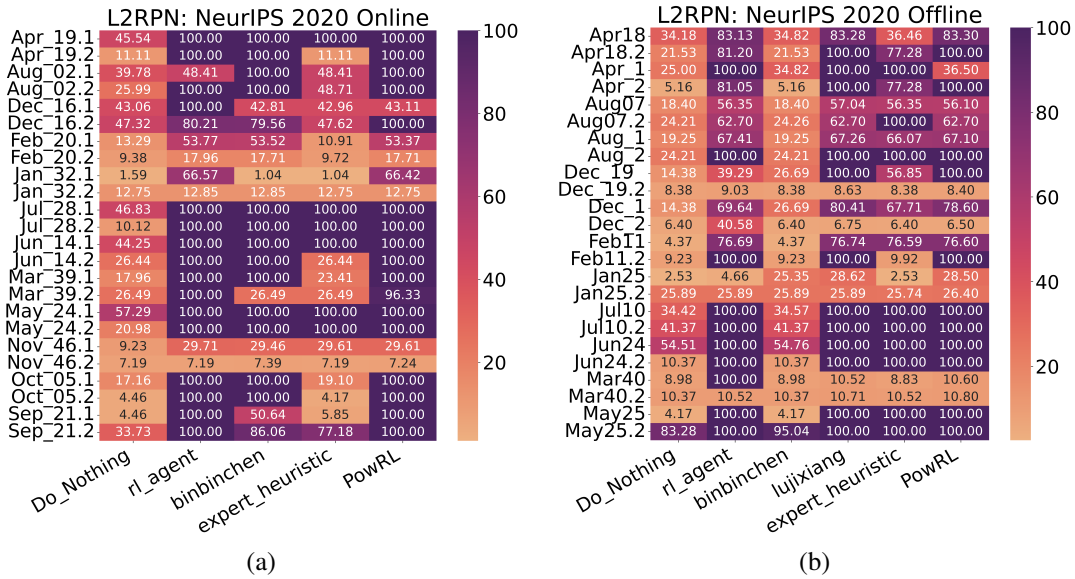
Figure 4: Survival percentages of various agents on the L2RPN NeurIPS 2020 challenge dataset (a) Online, (b) Offline.

bifurcated the network into two smaller groups). The critic network uses a single hidden layer (with ReLU activation) of dimension 64. The output dimension of the critic network is 1 and it uses the Gumbel-max trick to sample an action from the logits/probability distribution. The clipped PPO algorithm is implemented in PyTorch using Adam optimizer with a learning rate of 0.003, clip range of 0.2, rate/discount factor (γ) of 0.99, and generalized advantage estimation (λ) of 0.95. A single PPO agent was trained on the 36 parallel environments inside Grid2Op. The RL agent is engaged only when the maximum line overflow exceeds the safety threshold. Any such transition is stored in the prioritized replay buffer, and a long sequence of transitions is sampled (around 20k) in each epoch during the training.

Baseline algorithms We benchmark the PowRL framework against the other popular baseline approaches for power grid control. Readers are advised to refer to the section on Related Works for more details on the baseline algorithms.

1. Do_Nothing The most basic of all algorithms, it contains only a single action of `do-nothing`, which is passed on to the environment at every step for all the scenarios.

2. expert_heuristic This approach uses a combination of different actions collected based on the expert experience for the given power grid, where specific actions are taken in accordance with different situations faced.

3. lujixiang, the 3rd-place winner in NeurIPS 2020 challenge. This approach used 'guided exploration', where the top-5 Q-value actions were sampled and simulated, and the action with best simulated reward was selected.

4. binbinchen the 2nd-place winner in NeurIPS 2020 challenge. It uses an imitation learning framework, where a simple feed-forward neural network ('Junior' student) is trained to imitate an expert agent, while a PPO-based 'Senior' student is developed using the 'Junior' student as the Actor.

5. rl_agent, the winner of both the tracks in NeurIPS 2020 competition. It uses a black-box optimization approach to find optimal policies. In addition, for specific scenarios, such

as in the month of March, the agent uses additional generator redispatch guided by the prior experience to survive some critical attacks in those scenarios.

Results The aim of the RL agent is to explore the flexibilities offered by cheap topology actions meant to support human decision-making operations. The performance of the proposed PowRL agent with baselines is evaluated over the three different test datasets mentioned above. Each scenario is evaluated/scored based on the operation cost and the losses due to blackout, hence we have used both survival steps and scenario score as the evaluation criterion.

We first benchmark the PowRL on NeurIPS 2020 online test dataset which is 24 different weekly episodic scenarios. Fig. 4a reflects the scenario-wise performance of our approach along with the other baselines. Both the PowRL and the rl_agent fully survive the highest number of 16 scenarios, while the binbinchen and expert heuristics are able to survive fully only in 13 and 6 scenarios, respectively. It was found that a set of scenarios (January, February, November, December) looked tough, and required some new flexibilities, such as voltage control and load shedding, however, they were not made available in the environment. The dec_16.2 scenario has one of the strongest power grid attacks (Marot et al. 2021) and only our approach is able to survive successfully in this scenario. The other scenario is aug_02.1 where the PowRL and binbinchen successfully survive the full episode. The other difficult scenario is mar_39.2, where our agent survives nearly 96.3% only through topological actions. Though the rl_agent survives the full episode, it uses forced costly generator dispatch along with topology actions. Interestingly, while the rl_agent, and the binbinchen were the eventual winners in the competition (announced in Nov'20), our approach now sits at the top of the challenge leaderboard, outperforming all other agents (see Table. 1).

We next evaluate the different approaches on NeurIPS 2020 offline test dataset (see Fig. 4b). Interestingly, the lu-
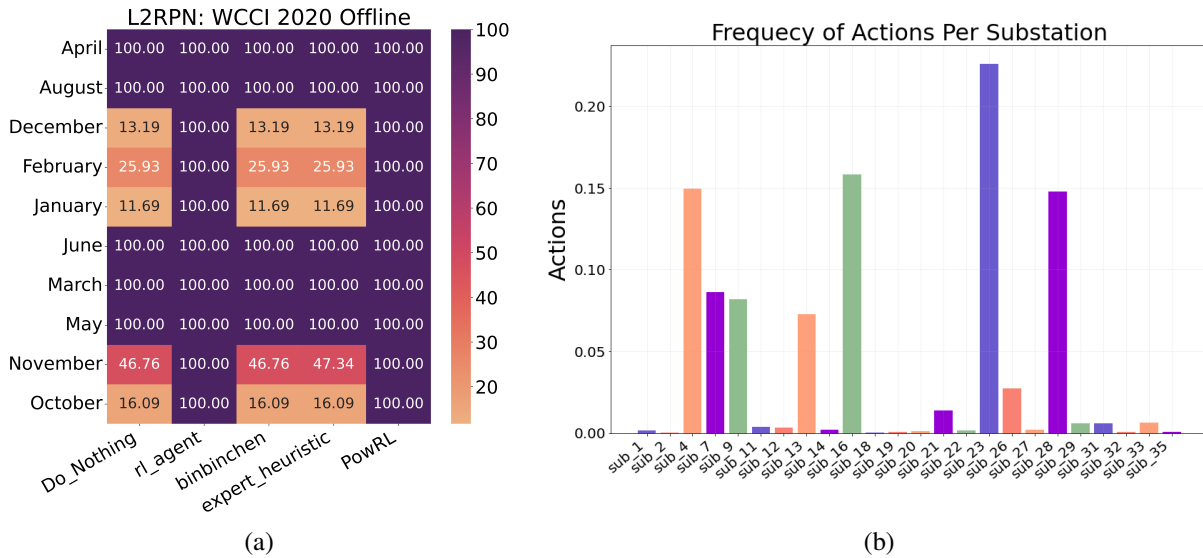
Figure 5: (a) Survival percentages of various agents on the L2RPN WCCI 2020 challenge (Offline). (b) Diversity of remedial actions distributed across multiple substations.

jixiang agent, which was the eventual 3rd place winner in the competition, survives the most 12 scenarios, while PowRL and the rl_agent survive 11 and 10 scenarios, respectively. Note that the source code included with the lujixiang appears broken and hence, a scenario-wise analysis for the online data could not be obtained. However, the offline scenario-wise analysis was obtained from their talk during the winner announcement. Despite surviving a fewer number of scenarios, PowRL still outperforms the lujixiang agent indicating that our model requires less operational cost and tries to avoid losses due to blackouts. The PowRL is able to survive the highest number of steps in the dec_2, feb11, and jan25.2 scenarios. In the Aug07.2 scenario, only the expert heuristic is able to survive the attack, as they are designed with the simple heuristic based on the generator dispatch and topological actions, rl_agent deploys additional heuristic generator redispatch, which results in the best performance in the march scenario in comparison to other algorithms. The scenario named dec_19 comprises of one of the strongest adversarial attacks and only PowRL survives it successfully. The binbinchen agent fares very poorly on this dataset and is unable to survive even a single scenario.

Finally, we evaluate the agents on the WCCI offline test dataset (see Fig. 5a). The PowRL agent and the rl_agent successfully survive all the scenarios present in this dataset, which is expected since this test dataset aims primarily with congestion management (no adversaries). Nonetheless, it is interesting to note that the performance of the binbinchen agent drops significantly when tested on this test dataset, too, which clearly indicates overfitting issues during its training. All the results are summarized in Table. 1. It is also worth noting that the PowRL agent uses the least number of actions compared to other learning based agents. Despite using a much smaller action set, it is still able to achieve state-of-the-art-performance on all datasets. Additionally, the run-time with PowRL agent evaluated on the NeurIPS Online

challenge is the least among all RL agents (see Table. 1).

Post-hoc analysis We also analyzed the action sequences utilized by the PowRL agent on the NeurIPS online dataset. Fig. 5b shows the frequency with which an action at a substation is utilized by the PowRL agent in response to different attacks. We found out that the PowRL uses a total of 146 different actions spread across 25 different substations, which is notably diverse than the agents binbinchen (20 subs), rl_agent (18 subs), and lujixiang (16 subs) (Marot et al. 2021). Evidently, the most frequent actionable substations are 4, 16, 23, and 28 because of the availability of large number of potential topological combinations. Additionally, the PowRL agent also engages substations 11, 18, 19, 20, and 22, though they have relatively fewer number of elements. Unlike the rl_agent that encodes fixed sequences of actions on three substations that need to be acted on during instances of heavy overload, our PPO framework is capable of learning the optimal sequence combinations during the training phase itself. For instance, our analysis finds out the most frequently occurring combinations of substations during events of critical attacks in January: $[9 \rightarrow 13 \rightarrow 23 \rightarrow 16]$, $[7 \rightarrow 16 \rightarrow 23 \rightarrow 13]$, $[23 \rightarrow 9 \rightarrow 7 \rightarrow 13 \rightarrow 16]$, and in December: $[21 \rightarrow 23 \rightarrow 16]$.

Conclusion

This study proposes a heuristic-guided RL framework, PowRL, for robust control of power networks subjected to production and demand uncertainty, as well as adversarial attacks. Using a careful action selection process, in combination with line reconnection and recovery heuristics, equips PowRL to outperform SOTA approaches on several challenge datasets even with reduced action space. PowRL not only diversifies its actions across substations, but also learns to identify important action sequences to protect the network against targeted adversarial attacks.

References

Alhazmi, M.; Dehghanian, P.; Wang, S.; and Shinde, B. 2019. Power grid optimal topology control considering correlations of system uncertainties. *IEEE Transactions on Industry Applications*, 55(6): 5594–5604.

Arwa, E. O.; and Folly, K. A. 2020. Reinforcement learning techniques for optimal power control in grid-connected microgrids: A comprehensive review. *IEEE Access*, 8: 208992–209007.

Baranwal, M.; Askarian, A.; Salapaka, S.; and Salapaka, M. 2018. A distributed architecture for robust and optimal control of DC microgrids. *IEEE Transactions on Industrial Electronics*, 66(4): 3082–3092.

Dehghanian, P.; Wang, Y.; Gurrala, G.; Moreno-Centeno, E.; and Kezunovic, M. 2015. Flexible implementation of power system corrective topology control. *Electric Power Systems Research*, 128: 79–89.

Donnot, B. 2020a. Grid2op- A testbed platform to model sequential decision making in power systems. . https://GitHub.com/rte-france/grid2op. Accessed: 2022-04-15.

Donnot, B. 2020b. L2RPN NEURIPS 2020 - Robustness Track. https://competitions.codalab.org/competitions/25426#learn_the_details-get_starting_kit. Accessed: 2022-04-15.

Donnot, B. 2020c. L2RPN WCCI 2020 Competition. https://competitions.codalab.org/competitions/24902. Accessed: 2022-01-15.

Donnot, B. 2020d. Lightsim2grid - A C++ backend targeting the Grid2Op platform. https://GitHub.com/bdonnot/lightsim2grid. Accessed: 2022-04-15.

Donnot, B.; Guyon, I.; Schoenauer, M.; Panciatici, P.; and Marot, A. 2017. Introducing machine learning for power system operation support. *arXiv preprint arXiv:1709.09527*.

Hu, Y.; Chen, B.; and Tang, K. 2021. L2RPN_NIPS_2020_a_PPO_Solution. . https://github.com/AsprinChina/L2RPN_NIPS_2020_a_PPO_Solution. Accessed: 2022-04-15.

Lan, T.; Duan, J.; Zhang, B.; Shi, D.; Wang, Z.; Diao, R.; and Zhang, X. 2020. AI-based autonomous line flow control via topology adjustment for maximizing time-series ATCs. In *2020 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5. IEEE.

Lee, K.; Lee, B.-U.; Shin, U.; and Kweon, I. S. 2020. An efficient asynchronous method for integrating evolutionary and gradient-based policy search. *Advances in Neural Information Processing Systems*, 33: 10124–10135.

Marot, A.; Donnot, B.; Dulac-Arnold, G.; Kelly, A.; O'Sullivan, A.; Viebahn, J.; Awad, M.; Guyon, I.; Panciatici, P.; and Romero, C. 2021. Learning to run a power network challenge: a retrospective analysis. In *NeurIPS 2020 Competition and Demonstration Track*, 112–132. PMLR.

Marot, A.; Donnot, B.; Romero, C.; Donon, B.; Lerousseau, M.; Veyrin-Forrer, L.; and Guyon, I. 2020. Learning to run a power network challenge for training topology controllers. *Electric Power Systems Research*, 189: 106635.

Marot, A.; Donnot, B.; Tazi, S.; and Panciatici, P. 2018. Expert system for topological remedial action discovery in smart grids. In *Mediterranean Conference on Power Generation, Transmission, Distribution and Energy Conversion (MEDPOWER 2018)*, 1–6. IET.

Ren, H.; Dobson, I.; and Carreras, B. A. 2008. Long-term effect of the n-1 criterion on cascading line outages in an evolving power transmission grid. *IEEE transactions on power systems*, 23(3): 1217–1225.

Schaul, T.; Quan, J.; Antonoglou, I.; and Silver, D. 2016. Prioritized Experience Replay. In *ICLR (Poster)*.

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

Wang, Q.; and McCalley, J. D. 2013. Risk and "N-1" criteria coordination for real-time operations. *IEEE Transactions on Power Systems*, 28(3): 3505–3506.

Wang, Z.; Schaul, T.; Hessel, M.; Hasselt, H.; Lanctot, M.; and Freitas, N. 2016. Dueling network architectures for deep reinforcement learning. In *International conference on machine learning*, 1995–2003. PMLR.

Wulf, W. A. 2000. Great achievements and grand challenges. *The Bridge*, 30(3): 4.

Yoon, D.; Hong, S.; Lee, B.-J.; and Kim, K.-E. 2021. Winning the L2{RPN} Challenge: Power Grid Management via Semi-Markov Afterstate Actor-Critic. In *International Conference on Learning Representations*.

Zhao, Y.; Yuan, C.; Liu, G.; and Grinberg, I. 2018. Graph-based Preconditioning Conjugate Gradient Algorithm for" N-1" Contingency Analysis. In *2018 IEEE Power & Energy Society General Meeting (PESGM)*, 1–5. IEEE.

Zhihong, Y.; Chunlei, X.; Jixiang, L.; Hongsheng, X.; Kang, X.; Tianyu, C.; Zhilin, X.; Junjun, L.; Jinjun, L.; and Tianhua, C. 2020. A winning approach of NeurIPS_2020_L2RPN_Comp. . https://github.com/lujasone/NeurIPS_2020_L2RPN_Comp_An_Approach.

Zhou, B.; Zeng, H.; Liu, Y.; Li, K.; Wang, F.; and Tian, H. 2021. Action set based policy optimization for safe power grid management. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*, 168–181. Springer.