Programmatic Reinforcement Learning for Trustworthy Microgrid Management

Subrat Prasad Panda [1,2] **Blaise Genest** [2,3] **Arvind Easwaran** [1]

[1]*College of Computing and Data Science (CCDS), NTU Singapore* [2]*CNRS@CREATE, Singapore* [3]*CNRS, IPAL, France.*
Correspondence to: Subrat Prasad Panda subratpr001@e.ntu.edu.sg.

1. Introduction

To meet rising energy demands while ensuring sustainability and reducing environmental impact, there is a major shift toward clean energy generated from renewable sources like solar and wind. These distributed energy resources are increasingly integrated into existing power grids, with microgrids enabling the local integration and consumption of renewable energy [1]. A microgrid typically includes photovoltaic (PV) panels, energy storage systems (ESS), and diesel generators (DG) and is often connected to the main grid for energy exchange. The Energy Management System (EMS) optimizes operational costs by scheduling power generation, storage, and distribution [2].

EMS faces challenges in managing microgrid energy due to uncertainties in renewable energy generation and load demand, compounded by the need to commit energy exchange plans with the main grid a day in advance [3, 4]. This is particularly difficult due to the long-horizon optimization. To address this, previous work has adopted hierarchical approaches in EMS [5, 4], splitting the problem into two stages and using optimization methods like Model Predictive Control (MPC) or learning-based Deep Reinforcement Learning (DRL). Although DRL adapts better to uncertain environments than MPC [2, 4], it relies on black-box neural network (NN) policies that hinder interpretability by energy scientists and do not meet the requirements for trustworthy policies in safety-critical energy systems [6, 7]. To address these challenges, programmatic policies, such as decision trees (DTs) or domain-specific programs have gained significant attention in DRL, as they can be easily formulated and understood by energy experts, and because they generate policies which are amenable to verification [8, 9], unlike NN policies. Additionally, their well-defined structure allows seamless injection of domain knowledge.

Learning programmatic policies within the DRL framework is however challenging due to their discrete, non-differentiable nature, which hinders gradient descent-based training. As a result, there is growing interest in Differentiable Programmatic Reinforcement Learning (∂PRL), with prior attempts such as [10, 11] that rely on smooth approximations, which impact performance. A recent work, DTSemNet [12], proposed a novel method to overcome these limitations, demonstrating the ability to learn hard DT policies with performance comparable to NN policies in benchmark DRL environments. The potential of ∂PRL in the energy management of microgrids, however, has not been explored before.

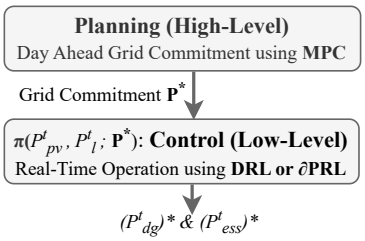

Fig. 1: Hierarchical operation in a microgrid.

This work explores application of ∂PRL, particularly DTSemNet as a potential replacement for NN-based policies in EMS. By leveraging ∂PRL, we aim to retain the uncertainty handling of DRL while ensuring transparency and verifiability of microgrid controllers. We integrate DTSemNet in the controller of a hierarchical operation framework in microgrid, as studied in [4], where a high-level planning agent plans energy commitments and a low-level controller agent optimizes real-time dispatch. DTSemNet, as the low-level controller, achieves performance comparable to NN-based controllers while ensuring a trustworthy microgrid EMS.

2. Hierarchical Operation in Microgrids

A grid-connected microgrid must establish a day-ahead energy exchange plan/commitment with the grid operator to ensure stability of the main grid. This energy exchange commitment, denoted as tuple $\mathbf{P} = (P_g^1, P_g^2, \ldots, P_g^T)$, represents the power exchange commitments for the next T time steps, where P_g^t is the scheduled power exchange at future time step t. At $t = 1$, the microgrid commits to this plan, which is reevaluated after T steps (e.g., at midnight). During real-time operation, the microgrid strictly follows the pre-determined grid exchange commitments while dynamically adjusting DG power setpoints (P_{dg}^t) and ESS setpoints (P_{ess}^t) based on actual solar generation (P_{pv}^t) and load demand (P_l^t) at each time step t. The goal is to determine the optimal day-ahead commitment, $\mathbf{P}^*$, and the optimal real-time operation setpoints for DG power, $(P_{dg}^t)^*$, and ESS power, $(P_{ess}^t)^*$, that minimizes operational cost. However, due to uncertainties in weather conditions and load demand, this long-horizon optimization problem is challenging to solve. To address this, we adopt a two-level hierarchical approach similar to [4]: the high-level *planner* agent optimizes the day-ahead grid commitment to determine $\mathbf{P}^*$, while the low-level *controller* agent ensures adherence to this plan and computes the optimal DG power setpoints, $(P_{dg}^t)^*$, and ESS power setpoints, $(P_{ess}^t)^*$, at each time step, as shown in Fig. 1.

The following describes these components:

Planning: The planning agent uses multi-scenario Stochastic Programming (SP), an MPC method that considers various solar generation and load demand scenarios. This ensures the microgrid meets commitments under different conditions and generates a commitment plan, $\mathbf{P}^*$, for the next T time steps. Here, we use hourly steps with $T = 24$, re-planning every 24 hours.

Control: With the commitment plan $\mathbf{P}^*$ set by the planning agent, the low-level controller ensures real-time adherence to it while adapting to deviations in solar generation P_{pv}^t and load demand P_l^t. The control policy $\pi(P_{pv}^t, P_l^t; \mathbf{P}^*)$ determines the DG power setpoint at each time step t. The policy is learned using the DRL method SAC [13], and it outputs a continuous DG setpoint, while the ESS setpoint is based on the remaining energy requirement.

3. ∂PRL for Real-Time Control

Programmatic policies are those that can be represented using a domain-specific language (DSL) [14], which is defined by energy experts. It is defined as a pair $(\mathcal{T}, \theta)$, where $\mathcal{T}$ specifies the discrete policy architecture, and θ represents continuous learable parameters [15]. The structure of $\mathcal{T}$ follows a context-free grammar based on the DSL, as shown in Fig. 2. Let $\mathcal{X} \in \mathbb{R}^n$ denote the input and $\mathcal{Y} \in \mathbb{R}^m$ the output to the programmatic policy, where n and m are the dimension on input and output space, respectively. The program expression is constructed using non-terminals $\mathcal{T}$ and $\mathcal{C}$, which define the computation. Here, the branch condition $\mathcal{C}$ and the leaf node expression $\mathcal{L}$ are defined as the differentiable functions $f_\theta(\mathcal{X}) \in \mathbb{R}$ and $g_\theta(\mathcal{X}) \in \mathbb{R}^m$, respectively. These functions are defined according to the requirements of the energy domain. The program expression $\mathcal{T}$ evaluates to an output (i.e., an action in DRL) for a given input (i.e., a state in DRL) with $\mathcal{Y} := [\![\mathcal{T}]\!](\mathcal{X})$.

$$\mathcal{T} ::= \mathcal{L} \mid \textbf{if } \mathcal{C} \textbf{ then } \mathcal{T} \textbf{ else } \mathcal{T}$$
$$\mathcal{C} ::= f_\theta(\mathcal{X}) \geq 0$$

Fig. 2: DSL Grammar for programmatic policies.

Learning programmatic policies is challenging because they do not support direct gradient-based optimization. Prior work has explored two main approaches. The first relies on imitation learning, where a NN is trained first, and a programmatic policy is then derived by imitating it [8]. For instance, in [16], a DT policy is learned by imitating the output of an MPC controller in a microgrid. However, this approach is inefficient, as the quality of the learned policy depends on dataset construction. The second approach, ∂PRL, aims to make programmatic policies differentiable and integrate them directly into the DRL framework, which has been shown to outperform imitation learning-based methods as demonstrated in [10, 12].

3.1 DT Policy as a Controller in Microgrids

When both f_θ and g_θ are affine, i.e., $:= \theta_b + \theta.\mathcal{X}$, the DSL definition in Fig. 2 corresponds to an oblique DT with a linear controller at each leaf, which can be used as a microgrid controller agent, $\pi(\cdot) := [\![\mathcal{T}]\!](\cdot)$. Traditionally, training such policies via gradient descent relied on approximations like sigmoid relaxations or the Straight-Through Estimator (STE). The work in [17] attempts to learn a soft DT as a microgrid controller, but hardening it introduces inaccuracies [10]. A recent state-of-the-art approach avoids these approximations by proposing Decision Tree Semantic Network (DTSemNet), a novel encoding method that represents DTs as NNs. DTSemNet has demonstrated strong performance across classification, regression, and DRL benchmarks. In this work, we leverage capability of DTSemNet to learn a low-level DT control policy in EMS, regulating DG power for real-time dispatch.

Policy	MPC Perfect	NN Policy	DTSemNet
Cost ($)	1809	1844	1846

Table 1: Operation cost by different policies.

3.2 Experiment and Results

We follow the same experimental setup and dataset as [4], which consists of real-world hourly load demand and solar generation data. Data from January to November is used for training, while December is reserved for testing. The training data is used to generate solar and demand profiles for SP-based optimization and to train the low-level NN and DTSemNet policies. The test data is used to evaluate the average hourly operation costs. For benchmarking, we use MPC Perfect, which utilizes actual future data (typically unavailable in real-world scenarios) to compute the optimal average hourly cost. As shown in Table 1, DTSemNet (height 8) performs comparably to the NN policy (128×128), with a similar number of trainable parameters and offering additional interpretability.

4. Conclusion and Future Work

This work explores the use of ∂PRL for trustworthy microgrid energy management, to address the interpretability and verifiability limitations of NN-based controllers in DRL frameworks. DTSemNet facilitates a transparent decision-making process that energy scientists may find easy to grasp, as its output can be written in a formalism close to their expertise while maintaining competitive performance with NNs. This should also enable scientists to incorporate their knowledge into the formalism (e.g. safety constraints [18]), which we will explore next. This advances our long-term goal, which is to produce scientific discovery directly through ∂PRL, where the causal relationship (e.g. physical equations) between various attributes/features is discovered through learning.

Acknowledgments

This research was conducted as part of the DesCartes program and was supported by the National Research Foundation, Prime Minister's Office, Singapore, under the Campus for Research Excellence and Technological Enterprise (CREATE) program. This research/project is also supported by the National Research Foundation, Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-017).

References

[1] Lei Chen, Lingyun Gao, Shuping Xing, Zhicong Chen, and Weiwei Wang. Zero-carbon microgrid: Real-world cases, trends, challenges, and future research prospects. *Renewable and Sustainable Energy Reviews*, 203:114720, 2024.

[2] Noer Fadzri Perdana Dinata, Makbul Anwari Muhammad Ramli, Muhammad Irfan Jambak, Muhammad Abu Bakar Sidik, and Mohammed M Alqahtani. Designing an optimal microgrid control system using deep reinforcement learning: A systematic review. *Engineering Science and Technology, an International Journal*, 51:101651, 2024.

[3] Dimple Raja Prathapaneni and Ketan P. Detroja. An integrated framework for optimal planning and operation schedule of microgrid under uncertainty. *Sustainable Energy, Grids and Networks*, 19:100232, 2019.

[4] Subrat Prasad Panda, Blaise Genest, Arvind Easwaran, Rémy Rigo-Mariani, and Pengfeng Lin. Methods for mitigating uncertainty in real-time operations of a connected microgrid. *Sustainable Energy, Grids and Networks*, 38:101334, 2024.

[5] Zitong Zhang, Jing Shi, Wangwang Yang, Zhaofang Song, Zexu Chen, and Dengquan Lin. Deep reinforcement learning based bi-layer optimal scheduling for microgrid considering flexible load control. *CSEE Journal of Power and Energy Systems*, pages 1–14, 2022.

[6] AI HLEG. Ethics guidelines for trustworthy AI. https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai, 2019. [Accessed 19-11-2024].

[7] Mengdi Xu, Zuxin Liu, Peide Huang, Wenhao Ding, Zhepeng Cen, Bo Li, and Ding Zhao. Trustworthy reinforcement learning against intrinsic vulnerabilities: Robustness, safety, and generalizability. *arXiv preprint arXiv:2209.08025*, 2022.

[8] Osbert Bastani, Yewen Pu, and Armando Solar-Lezama. Verifiable reinforcement learning via policy extraction. *Advances in neural information processing systems*, 31, 2018.

[9] Abhinav Verma, Hoang Le, Yisong Yue, and Swarat Chaudhuri. Imitation-projected programmatic reinforcement learning. *Advances in Neural Information Processing Systems*, 32, 2019.

[10] Rohan R. Paleja, Yaru Niu, Andrew Silva, Chace Ritchie, Sugju Choi, and Matthew C. Gombolay. Learning interpretable, high-performing policies for continuous control problems. In *Robotics: Science and Systems*, 2022.

[11] Ajaykrishna Karthikeyan, Naman Jain, Nagarajan Natarajan, and Prateek Jain. Learning accurate decision trees with bandit feedback via quantized gradient descent. *Transactions on Machine Learning Research*, 2022.

[12] Subrat Prasad Panda, Blaise Genest, Arvind Easwaran, and Ponnuthurai Nagaratnam Suganthan. *Vanilla Gradient Descent for Oblique Decision Trees*. IOS Press, October 2024.

[13] Tuomas Haarnoja, Aurick Zhou, Kristian Hartikainen, George Tucker, Sehoon Ha, Jie Tan, Vikash Kumar, Henry Zhu, Abhishek Gupta, Pieter Abbeel, et al. Soft actor-critic algorithms and applications. *arXiv preprint arXiv:1812.05905*, 2018.

[14] John E. Hopcroft, Rajeev Motwani, and Jeffrey D. Ullman. *Introduction to Automata Theory, Languages, and Computation*. Addison-Wesley, 3rd edition, 2007. Pearson International Edition.

[15] Wenjie Qiu and He Zhu. Programmatic reinforcement learning without oracles. In *International Conference on Learning Representations*, 2022.

[16] Yuchong Huo, François Bouffard, and Géza Joós. Decision tree-based optimization for flexibility management for sustainable energy microgrids. *Applied Energy*, 290:116772, 2021.

[17] Gargya Gokhale, Seyed Soroush Karimi Madahi, Bert Claessens, and Chris Develder. Distill2explain: Differentiable decision trees for explainable reinforcement learning in energy application controllers. In *Proceedings of the 15th ACM International Conference on Future and Sustainable Energy Systems*, e-Energy '24, page 55–64, New York, NY, USA, 2024. Association for Computing Machinery.

[18] Chenxi Yang and Swarat Chaudhuri. Safe neurosymbolic learning with differentiable symbolic execution. *arXiv preprint arXiv:2203.07671*, 2022.