PROCEEDINGS OF ECOS 2023 - THE 36TH INTERNATIONAL CONFERENCE ON
**E**FFICIENCY, **C**OST, **O**PTIMIZATION, **S**IMULATION AND ENVIRONMENTAL IMPACT OF ENERGY SYSTEMS
25-30 JUNE 2023, LAS PALMAS DE GRAN CANARIA, SPAIN

# Reinforcement Learning for Joint Design and Control of Battery-PV Systems

***Marine Cauz[a,b], Adrien Bolland[c], Bardhyl Miftari[c], Lionel Perret[b], Christophe Ballif[a,d], and Nicolas Wyrsch[a]***

*[a] École polytechnique fédérale de Lausanne (EPFL), Institute of Electrical and Micro Engineering (IEM), Photovoltaics and Thin-Film Electronics Laboratory (PV-Lab), Neuchâtel, Switzerland. Marine.Cauz@epfl.ch, CA.*

*[b] Planair SA, Yverdon-les-bains, Switzerland*

*[c] Department of Electrical Engineering and Computer Science, University of Liège, Liège, Belgium*

*[d] Centre Suisse d'Electronique et de Microtechnique (CSEM), PV-Center, Neuchâtel, Switzerland*

**Abstract:**

The decentralisation and unpredictability of new renewable energy sources require rethinking our energy system. Data-driven approaches, such as reinforcement learning (RL), have emerged as new control strategies for operating these systems, but they have not yet been applied to system design. This paper aims to bridge this gap by studying the use of an RL-based method for joint design and control of a real-world PV and battery system. The design problem is first formulated as a mixed-integer linear programming problem (MILP). The optimal MILP solution is then used to evaluate the performance of an RL agent trained in a surrogate environment designed for applying an existing data-driven algorithm. The main difference between the two models lies in their optimization approaches: while MILP finds a solution that minimizes the total costs for a one-year operation given the deterministic historical data, RL is a stochastic method that searches for an optimal strategy over one week of data on expectation over all weeks in the historical dataset. Both methods were applied on a toy example using one-week data and on a case study using one-year data. In both cases, models were found to converge to similar control solutions, but their investment decisions differed. Overall, these outcomes are an initial step illustrating benefits and challenges of using RL for the joint design and control of energy systems.

**Keywords:**

Energy systems, Design, Control, RL, MILP.

## 1. Introduction

### 1.1. Background and related work

The current transition to renewable energy sources requires rethinking new energy systems, characterized by decentralized and intermittent production. The development of these systems typically occurs in two distinct steps, namely the design and control of these systems. The design problem involves identifying the design variables which are the optimal size of energy system components. The control problem aims to determine the control variables which are the optimal actions to operate the energy system components. Both design and control problem should jointly minimize a cost function and are typically solved sequentially. This paper explores the value of solving the design and control tasks, using a reinforcement learning (RL) method as appropriate design is intrinsically linked to subsequent operation. To evaluate the effectiveness of this approach, its performance are compared with that of the Mixed Integer Linear Programming (MILP) method.

On the one hand, RL is a data-driven approach where an agent learns to make decisions in a dynamic environment through trial-and-error experience. It involves an agent interacting with an environment and receiving feedback in the form of rewards or penalties based on its actions, with the goal of maximizing its cumulative reward over time. On the other hand, Mixed Integer Linear Programming (MILP) is a mathematical optimization technique used to solve problems with linear constraints and integer variables. It involves formulating a mathematical model of the problem and using an optimization algorithm to find the best solution. Both RL and MILP methods will be used to benchmark the results of a one-year time series.

As highlighted in a recent review [1], RL-based approaches have significant potential, yet not fully exploited, in the energy field. Specifically, the review points out that energy systems are typically designed using either MILP or heuristic methods, with RL approaches dedicated to their control. Integrating RL beyond energy flow control would open new interesting research questions. In [2], RL is used to support distributed energy system design due to its flexibility and model-free nature, which allows it to be adapted to different environments at different

scales. However, they did not simultaneously address the dispatch and design problem as a distributed reward problem, as done in this work. Instead, they used a cooperative coevolution algorithm (COCE) to assist the optimization process. Jointly addressing the design and operation of energy systems is a key issue, especially for multi-energy systems, as discussed in [3], where multi-objective evolutionary algorithms (EMOO) and MILP are used to integrate biomass technologies in a multi-energy system. In [4], the focus is on evolution algorithms and their comparison with deep reinforcement learning strategies. After clarifying the fundamental differences between the two approaches, the discussion revolves around their ability to parallelize computations, explore environments, and learn in dynamic settings. The potential of hybrid algorithms combining the two techniques is also investigated, along with their real-world applications.

RL-based frameworks are successfully applied to the operation of energy systems [5], although these methods have not, to the authors' knowledge, been extended to solve real-world design problem in energy system. As reviewed in [6], RL-based frameworks are popular for addressing electric vehicle (EV) charging management, mostly with variants of the DQN algorithm, and outperform other traditional methods. In [7], various deep RL algorithms are benchmarked against rule-based control, model predictive control, and deterministic optimization in the presence of PV generation. The study, which aims to increase PV self-consumption and state-of-charge at departure, demonstrates the potential of RL for real-time implementation. For solving V2G control under price uncertainty, [8] modeled the problem with a Markov Decision Process (MDP) [9], a mathematical framework for modeling system where stochasticity is involved. Additionally, a linear MDP formulation is also used in [10] to address the coordination of multiple charging points at once. Finally in [11], a data-driven approach is defined and evaluated for coordinating the charging schedules of multiple EVs using batch reinforcement learning with a real use case. In conclusion, these studies provide valuable insights and tools for optimizing and improving energy systems, demonstrating the potential of RL to tackle the operation of complex energy systems.

### 1.2. Contribution

This work aims to evaluate the relevance of jointly designing and controlling an energy system using a deep RL approach. To achieve this purpose, two methods are benchmarked to address jointly the design and control problem of a real-world PV-battery system. The first method, MILP, computes the optimal design and control solution over a sequence of historical data. The second method, RL, computes the optimal design and a control policy through interactions with a simulator by trial and error. The specific RL algorithm used in this study is referred to as Direct Environment and Policy Search (DEPS) [12]. DEPS extends the REINFORCE algorithm [13] by combining policy gradient with model-based optimization techniques to parameterize the design variables. In this framework, an agent looks for the design and control variables that jointly maximize the expected sum of rewards collected over the time horizon of interest. The outcomes of both methods are discussed in the subsequent sections of this paper.

This paper is structured as follows. Section 2. provides two formulations of the energy system, one designed for MILP and the other for RL, and discusses the methodology used to benchmark the results. In Section 3., the outcomes of the study are presented, and these results are discussed in Section 4., with a focus on the potential of RL for joint design and control of energy systems. Finally, the paper concludes with a summary in Section 5..

## 2. Method

### 2.1. Problem statement

The study is carried out for the energy system illustrated in Figure 1, whose components are detailed in the subsections below. Overall, the system refers to an office building that has been fitted with a PV installation and a stationary lithium-ion battery to meet its own electricity consumption. Additionally, the building is connected to the electricity grid.

The objective of the study is to jointly propose a design of the PV and battery components, as well as a control strategy of the described energy system in order to minimize the total cost of its ownership. In the following Subsection 2.2., the system is expressed as a mathematical program made-up of constraints and objectives. To be more precise, it is tackled as a Mixed-Integer Linear Program. Subsection 2.3. formulates a surrogate environment as an MDP. The latter represents the same dynamics and rewards as the original problem but the objective is to maximize the sum of rewards gathered over one week on expectation over the 52 weeks of the year of data. By doing so, it allows the use of the RL algorithm and expects the optimal solution to be close to the solution of the original problem. Results are discussed in Section 3.. Finally, for both methods, the energy system is studied over a finite time horizon $T$, on which all costs are evenly distributed across each time step $t$. The methodology and the context of the experiments conducted are specified in Subsection 2.4..

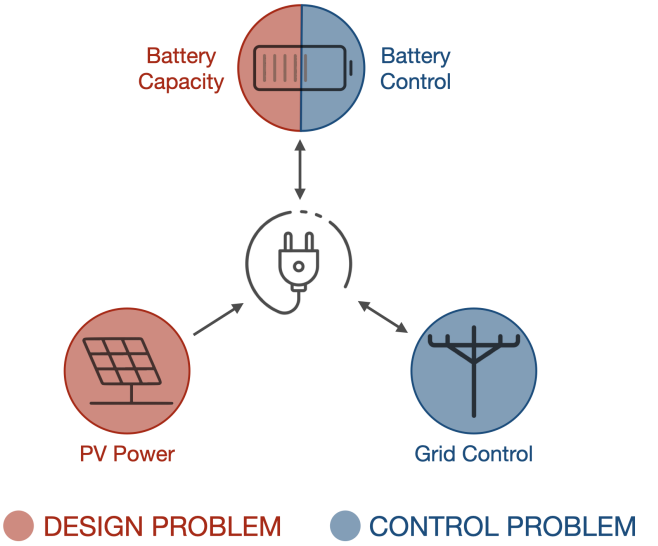


**Figure 1**: The energy system to be jointly designed and controlled is characterized by an electrical consumption, a battery, a photovoltaic system, and a grid connection. The design problem consists of determining the photovoltaic power and battery capacity, while the control problem aims to regulate the charging and discharging of the battery, as well as the import (resp. export) of electricity to (resp. from) the grid. The overall objective is to meet the electrical consumption needs while minimizing the costs of installing and operating the system.

## 2.2. Energy system

This subsection describes the physical constraints that apply to the components of the energy system. Theses components, in sequential order, consist of the PV panels, the battery, the electrical load and the power grid. The set of design and control variables and the parameters of the whole system, which is modeled as a discrete-time system, are gathered respectively in Table 1 and 2, respectively.

| | Variable | Set | Unit | Description |
|---|---|---|---|---|
| GRID | $P^{\text{IMP}}$ | $\mathbb{R}_+^T$ | kW | imported power (from the grid) |
| | $P^{\text{EXP}}$ | $\mathbb{R}_+^T$ | kW | exported power (to the grid) |
| PV | $P^{\text{NOM}}$ | $\mathbb{R}_+$ | $\text{kW}_p$ | nominal power of the PV installation |
| BATTERY | $B$ | $\mathbb{R}_+$ | kWh | nominal capacity of the battery |
| | SOC | $\mathbb{R}_+^T$ | kWh | state of charge of the battery |
| | $P^B$ | $\mathbb{R}^T$ | kW | power exchanged with the battery |

**Table 1**: Set of design and control variables of the energy system studied.

### PV system

The objective of the PV installation is to generate electricity on-site to fulfill the local electricity demand. The design of this component is one of the two design variables that will result from the optimization process. The range of the suitable nominal power $P^{\text{NOM}}$, corresponding to its design variable, is set in Eq. (1) and the production at time $t$ is directly proportional to this nominal design variable as shown in Eq. (2). The normalized annual curve $\overline{p}_t^{\text{PROD}}$ corresponds to the actual hourly averaged PV production power of the building.

$$P_{\text{MIN}}^{\text{NOM}} \leq P^{\text{NOM}} \leq P_{\text{MAX}}^{\text{NOM}} \tag{1}$$

$$P_t^{\text{PROD}} = P^{\text{NOM}} \cdot \overline{p}_t^{\text{PROD}} \tag{2}$$

The CAPEX and OPEX values, which are respectively the initial investment and the annual maintenance cost, of the installation are made up of a fixed and a variable part to take account of potential scale effects.

$$\text{CX}_{\text{PV}} = \text{CX}_{\text{PV}}^{\text{FIX}} + \text{CX}_{\text{PV}}^{\text{VAR}} \cdot P^{\text{NOM}} \tag{3}$$

$$\text{OX}_{\text{PV}} = \text{OX}_{\text{PV}}^{\text{FIX}} + \text{OX}_{\text{PV}}^{\text{VAR}} \cdot P^{\text{NOM}} \tag{4}$$

| | Parameter | Value | Set | Unit | Description |
|---|---|---|---|---|---|
| GRID | $C^{\text{IMP}}_{\text{GRID}}$ | 1 | $\mathbb{R}$ | €/kWh | imported electricity price |
| | $C^{\text{EXP}}_{\text{GRID}}$ | -0.05 | $\mathbb{R}$ | €/kWh | exported electricity price |
| | $C_{\text{GRID}}$ | | $\mathbb{R}^T$ | € | electricity network cost |
| | $P^{\text{MAX}}_{\text{GRID}}$ | 10'000 | $\mathbb{R}_+$ | kW | grid connection power |
| PV | $P^{\text{NOM}}_{\text{MIN}}$ | 0 | $\mathbb{R}_+$ | $\text{kW}_p$ | minimal nominal PV power |
| | $P^{\text{NOM}}_{\text{MAX}}$ | 200 | $\mathbb{R}_+$ | $\text{kW}_p$ | maximal nominal PV power |
| | $P^{\text{PROD}}$ | | $\mathbb{R}_+^T$ | kW | generated PV power |
| | $\overline{P}^{\text{PROD}}$ | | $\mathbb{R}_+^T$ | kW | expected generated PV power |
| | $\overline{p}^{\text{PROD}}$ | | $\mathbb{R}_+^T$ | kW | normalised PV power |
| | $L^{\text{PV}}$ | 20 | $\mathbb{N}$ | years | PV lifetime |
| | $\text{R}_{\text{PV}}$ | | $\mathbb{R}_+$ | - | annuity factor |
| | $\text{OX}^{\text{FIX}}_{\text{PV}}$ | 3 | $\mathbb{R}_+$ | € | OPEX PV fixed cost |
| | $\text{OX}^{\text{VAR}}_{\text{PV}}$ | 10 | $\mathbb{R}_+$ | €/kW | OPEX PV variable cost |
| | $\text{CX}^{\text{FIX}}_{\text{PV}}$ | 50 | $\mathbb{R}_+$ | € | CAPEX PV fixed cost |
| | $\text{CX}^{\text{VAR}}_{\text{PV}}$ | 200 | $\mathbb{R}_+$ | €/kW | CAPEX PV variable cost |
| BATTERY | $B_{\text{MIN}}$ | 0 | $\mathbb{R}_+$ | kWh | minimal nominal battery capacity |
| | $B_{\text{MAX}}$ | 200 | $\mathbb{R}_+$ | kWh | maximal nominal battery capacity |
| | $\eta^{\text{B}}$ | 0.9 | $]0, 1]$ | - | battery efficiency |
| | $L^{\text{B}}$ | 30 | $\mathbb{N}$ | years | battery lifetime |
| | $\text{R}_{\text{B}}$ | | $\mathbb{R}_+$ | - | annuity factor |
| | $\text{OX}^{\text{FIX}}_{\text{B}}$ | 5 | $\mathbb{R}_+$ | € | OPEX Battery fixed cost |
| | $\text{OX}^{\text{VAR}}_{\text{B}}$ | 6 | $\mathbb{R}_+$ | €/kW | OPEX Battery variable cost |
| | $\text{CX}^{\text{FIX}}_{\text{B}}$ | 30 | $\mathbb{R}_+$ | € | CAPEX Battery fixed cost |
| | $\text{CX}^{\text{VAR}}_{\text{B}}$ | 110 | $\mathbb{R}_+$ | €/kW | CAPEX Battery variable cost |
| SYSTEM | $T$ | | $\mathbb{N}$ | - | time horizon |
| | $\Delta t$ | 1 | $\mathbb{R}_+^T$ | h | time steps |
| | $h_t$ | | $[0 : 23]$ | h | hour of the time step |
| | $r$ | 0.05 | $\mathbb{R}$ | - | discount rate |
| | $P^{\text{LOAD}}$ | | $\mathbb{R}_+^T$ | kW | uncontrollable electricity consumption |
| | $\overline{P}^{\text{LOAD}}$ | | $\mathbb{R}_+^T$ | kW | expected electricity consumption |

**Table 2**: Set of parameters of the energy system studied.

### Battery

To maximize the potential for on-site self-consumption, a stationary lithium-ion battery is available. The design of this component, corresponding to its capacity $B$, is the second design variable to determine during the optimization process. This battery capacity can vary in the range of Eq. (5).

$$B_{\text{MIN}} \leq B \leq B_{\text{MAX}} \tag{5}$$

The state of charge variable, $\text{SOC}_t$, changes as a function of the power exchanged with the battery denoted $P^B_t$. This power is constrained, for charging, by the nominal capacity, Eq. (6), and, for discharging, by the energy stored, Eq. (7). Additionally, the battery efficiency, denoted $\eta^{\text{B}}$, is assumed identical for both the charging and the discharging processes.

$$P^B_t \leq \frac{\text{B} - \text{SOC}_t}{\Delta t} \qquad \text{if } P^B_t \geq 0 \tag{6}$$

$$P^B_t \geq \frac{-\text{SOC}_t}{\Delta t} \qquad \text{if } P^B_t \leq 0 \tag{7}$$

Knowing the power exchanged with the battery, the state of charge can be updated:

$$\text{SOC}_{t+1} = \begin{cases} \text{SOC}_t + P^B_t \cdot \eta^{\text{B}} \cdot \Delta t & \text{if } P^B_t \geq 0 \\ \text{SOC}_t + \frac{P^B_t}{\eta^{\text{B}}} \cdot \Delta t & \text{if } P^B_t < 0 \end{cases} \tag{8}$$

At the beginning of the optimization, i.e., $t = 0$, the battery SOC is set to half of its capacity value, to initialize the model. Moreover, to avoid any artificial benefit, the final SOC is constrained to be equal to the initial value,

as formulated in Eq. (10).

$$\mathrm{SOC}_{t=0} = \frac{B}{2} \tag{9}$$

$$\mathrm{SOC}_{t=0} = \mathrm{SOC}_{t=T} \tag{10}$$

Similar to the PV plant, the CAPEX and OPEX of the battery consist of both fixed and variable parts.

$$\mathrm{CX_B} = \mathrm{CX_B^{FIX}} + \mathrm{CX_B^{VAR}} \cdot \mathrm{B} \tag{11}$$

$$\mathrm{OX_B} = \mathrm{OX_B^{FIX}} + \mathrm{OX_B^{VAR}} \cdot \mathrm{B} \tag{12}$$

**Electrical load**

The electrical load used in this project is real data from an office building in Switzerland. This consumption is monitored on an hourly basis and reflects the consumption patterns of office days. This building load power, $P_t^{\mathrm{LOAD}}$, is provided as input and corresponds to an actual measurement sampled by hours over a year.

**Electrical grid**

To absorb excess solar production or to meet the electricity consumption in the absence of local production, the system is connected to the low-voltage electrical grid. This connection is modeled here as a single balance equation, called the conservation of electrical power, shown in Eq. (13). The power imported from the grid is referred to as $P_t^{\mathrm{IMP}}$ and the power injected is referred to as $P_t^{\mathrm{EXP}}$.

$$P_t^{\mathrm{PROD}} + P_t^{\mathrm{IMP}} = P_t^{\mathrm{LOAD}} + P_t^{\mathrm{B}} + P_t^{\mathrm{EXP}} \tag{13}$$

The grid power value at each time $t$ is derived from Eq. 13, and the power limit can be described as follows.

$$0 \leq P_t^{\mathrm{IMP}} \leq P_{\mathrm{GRID}}^{\mathrm{MAX}} \tag{14}$$

$$0 \leq P_t^{\mathrm{EXP}} \leq P_{\mathrm{GRID}}^{\mathrm{MAX}} \tag{15}$$

Based on the import and export power, the total cost of supplying electricity through the network $C_{\mathrm{GRID}}$ can be computed.

$$C_{\mathrm{GRID}} = \sum_{t=0}^{T-1} C_{\mathrm{GRID},t} = \sum_{t=0}^{T-1} P_t^{\mathrm{IMP}} \cdot C_{\mathrm{GRID},t}^{\mathrm{IMP}} - P_t^{\mathrm{EXP}} \cdot C_{\mathrm{GRID},t}^{\mathrm{EXP}} \tag{16}$$

**Objective function**

The objective of this study is to propose a design for the PV and battery components, along with their dispatch, with the aim of minimizing the total cost of ownership. This objective function, of minimizing the overall cost of the system, can be formulated as follows.

$$\min \quad \mathrm{TOTEX} \tag{17}$$

The total cost of the system, denoted TOTEX, is composed of the CAPEX and OPEX of both PV and battery components, as well as the grid cost.

$$\mathrm{TOTEX} = \mathrm{OPEX} + \mathrm{CAPEX} + C_{\mathrm{GRID}} \tag{18}$$

$$\mathrm{OPEX} = \mathrm{OX_{pv}} + \mathrm{OX_B} \tag{19}$$

$$\mathrm{CAPEX} = \mathrm{CX_{pv}} \cdot R_{pv} + \mathrm{CX_B} \cdot R_B \tag{20}$$

The OPEX and grid cost are computed over a finite time period $T$. However, the CAPEX is an investment cost that is independent of $T$. To enable the adaptation of the investment cost to the project duration, an annuity factor $R$ adjusts the CAPEX for the finite time horizon $T$. This annuity factor is computed according to Eq. (21), by taking into account the values of $T$, the annual discount rate $r$, and the lifetime $L$ of the component. This formula includes a scaling factor $\frac{T}{8760}$ to adapt $R$ to the period $T$, based on the assumption that $T$ is expressed in hours since 8760 is the number of hours in a year.

$$R = \frac{r \cdot (1+r)^L}{(1+r)^L - 1} \cdot \frac{T}{8760} \tag{21}$$

## 2.3. MDP formulation

This section presents an alternative formulation of the problem as a Markov Decision Process (MDP), which is a well-established framework for modeling sequential decision-making problems. This alternative formulation is required for applying DEPS. More precisely, an MDP($S, A, P, R, T$), as presented below, consists of the following elements: a finite set of states $S$, a finite set of actions $A$, a transition function $P$, a rewards function $R$, and a finite time horizon $T$.

**State Space**

The state of the system can be fully described by

$$s_t = (h_t, d_t, \text{SOC}_t, \overline{P}_t^{\text{PROD}}, \overline{P}_t^{\text{LOAD}}) \tag{22}$$

$$\in S = \{0, ..., 23\} \times \{0, ..., 364\} \times [0, B] \times \mathbb{R}_+ \times \mathbb{R}_+ \tag{23}$$

- $h_t \in \{0, ..., 23\}$ denotes the hour of the day at time $t$. The initial value is set to 0.
- $d_t \in \{0, ..., 364\}$ denotes the day of the year at time $t$. The initial value is set randomly.
- $\text{SOC}_t$ is the state of charge of the battery at time $t$, this value is upper bounded by the nominal capacity of the installed battery B. The value is set initially to a random value during the training process and to half of its capacity during the validation process.
- $\overline{P}_t^{\text{PROD}}$ represents the expected PV power at time $t$. This value is obtained by scaling normalized historical data $\overline{p}_t^{\text{PROD}}$ with the total installed PV power ($P^{\text{NOM}}$) and considering $h_t$ and $d_t$ values.
- $\overline{P}_t^{\text{LOAD}}$ denotes the expected value of the electrical load at time $t$. The load profile is determined using historical data that corresponds to the same hour and day as the PV power.

**Action Space**

The action of the system corresponds to the power exchanged with the battery.

$$\widetilde{a}_t = (\widetilde{P}_t^B) \tag{24}$$

After projecting the action to fall within the acceptable range specified by Eq. (6) and (7), the resulting value is used as $a_t$, as shown in Eq. (25). This corresponds to the power exchanged with the battery, denoted $P_t^B$, this value is positive when the battery is being charged and negative when it is being discharged.

$$P_t^B = \begin{cases} \frac{\text{B} - \text{SOC}_t}{\Delta t} & \text{if } \widetilde{P}_t^B > \frac{\text{B} - \text{SOC}_t}{\Delta t} \\ \frac{\text{SOC}_t}{\Delta t} & \text{if } \widetilde{P}_t^B < -\frac{\text{SOC}_t}{\Delta t} \\ P_t^B & \text{otherwise} \end{cases} \tag{25}$$

**Transition Function**

Each time step $t$ in the system corresponds to one hour, which implies the evolution specified in Eq. (26) of the state variable $h$ and every 24 time steps, the day is incremented by 1.

$$h_{t+1} = (h_t + 1) \bmod 24 \tag{26}$$

$$d_{t+1} = \text{Int}(\frac{h_t + 1}{24}) \tag{27}$$

where the function *Int* takes the integer value of the expression in Eq. (27).

The $\text{SOC}_t$ of the battery is updated as Eq. (8), based on the projected action value, and all other state variables are taken from input data.

$$\overline{P}_{t+1}^{\text{PROD}} = \overline{p}_{h_{t+1}, d_{t+1}}^{\text{PROD}} \cdot P^{\text{NOM}} \tag{28}$$

$$\overline{P}_{t+1}^{\text{LOAD}} = \overline{p}_{h_{t+1}, d_{t+1}}^{\text{LOAD}} \tag{29}$$

**Reward Function**

The reward signal to optimize the agent's actions in RL serves a similar aim as the objective function in the MILP formulation. Therefore, the reward here is the opposite value of the TOTEX defined at Eq. (18). This cost is composed of (i) the investment cost, (ii) the operating cost and (iii) the cost from the purchase and resale of electricity from the grid defined in Eq. (16).

$$r_t = -\text{TOTEX}_t \tag{30}$$

$$= -\text{CAPEX} - \text{OPEX} - C_{\text{GRID},t} \tag{31}$$

$$= -\text{CAPEX} - \text{OPEX} - P_t^{\text{IMP}} \cdot C_{\text{GRID},t}^{\text{IMP}} + P_t^{\text{EXP}} \cdot C_{\text{GRID},t}^{\text{EXP}} \tag{32}$$

where the grid cost is the only time-dependent factor, while CAPEX and OPEX are fixed values for a specific value of $P^{\text{NOM}}$ and $B$.

## 2.4. Methodology

This subsection discusses the fundamental differences between the two methods (i.e., MILP and RL), along with the experimental protocol that was employed to compare the results. As discussed briefly earlier, although the same problem is aimed to be solved, the methods under study are fundamentally different.

MILP is a method for solving problems that involves optimizing a linear function of variables that are either integer or constrained by linear equalities, as the problem described in Subsection 2.2. The MILP algorithm solves the optimization problem by iteratively adjusting the values of the design and control variables, subject to the constraints, until it finds the optimal solution that maximizes or minimizes the objective function, depending on the problem's goal. This method is applied to the problem described in Subsection 2.2. over a one-year time horizon ($T$ = 8760). The solution is said to be computed with perfect foresight meaning that all variables are selected accounting for the future realization of (normally unknown) events in the time series, providing an optimistic upper bound on the true performance of the control and design. Concretely, the MILP problem is here encoded in the Graph-Based Optimization Modelling Language (GBOML) [14] paired with the Gurobi solver [15].

In contrast, RL is a stochastic optimization method that learns from experience through trials and errors. In this study, we use DEPS [12], an algorithm optimizing design and control variables in an MDP, as the one described in Subsection 2.3., with a finite-time horizon. The agent receives feedback in the form of rewards when it selects a particular design and performs specific actions. The objective of the agent is to maximize the expected cumulative reward, which drives it to learn a design and a control policy. Ideally, as with MILP, the time horizon should be annual, or cover the entire lifetime of the system, taking into account seasonal production and consumption fluctuations and/or equipment aging. However, such extended time horizons are unsuitable for this RL approach. Therefore, to strike a balance between a horizon that is short enough for DEPS and long enough to observe the consequences of decisions on the system, a horizon of $T$ = 168 hours, i.e., 7 days, is defined. Additionally, for each simulation, the initial day is sampled uniformly from the year-long data set and the initial state-of-charge of the battery is also sampled uniformly at random. As the reward is optimized on expectation over all days, the resulting design and control policy is expected to account for the seasonality and other different hazard in the historical data. The DEPS algorithm is trained on a predetermined number of iterations. The PV power and battery capacity values obtained from the last iteration of the algorithm are then taken as the values of the design variables and the final policy is used for the control.

Unlike MILP, the RL method does not secure optimality, therefore the experimental protocol aims to compare both results to see how far the RL solution is from the optimal one. The experimental protocol is conducted in two distinct scenarios to differentiate the impact of adding the design variables in the joint problem. The first control-only scenario (CTR) assessed the control variables when the design variables are fixed. The second scenario, considering both the control and design (CTR & DES) problem, allows for flexibility in designing the battery capacity and PV power, the two design variables. To benchmark the performance of both methods in each scenario (i.e., CTR and CTR & DES), the reward and income value are reported. The reward value is computing according to Eq. (32) for the RL method. To estimate the average reward value for the MILP method, all reward values $r_t$ are averaged over time horizons of $T$ = 168. Comparing the average cumulative reward value of the MILP method to that of the RL method provides a first benchmark for evaluating the performance of both approaches. However, as shown in Eq. (32), only the grid cost is time-dependent, while the CAPEX and OPEX depend solely on investment decisions. Therefore, the income value is defined as the average reward value, but it only includes the grid cost and can be computed as follows:

$$Income = \sum_{t=0}^{T-1} -P_t^{\text{IMP}} \cdot C_{\text{GRID},t}^{\text{IMP}} + P_t^{\text{EXP}} \cdot C_{\text{GRID},t}^{\text{EXP}} \tag{33}$$

Finally, the experiments are performed in two steps. First, to perform a simple comparative study, working on a same finite time horizon $T$ = 168, both methods are conducted using data from a single summer week. Second, the data set is extended to include the one-year data set.

# 3. Results

The energy system presented in Section 2. is solved using the RL and MILP approaches with parameter values listed in Table 2. To differentiate the performance of the DEPS algorithm for control and design aspects, the study is conducted in two distinct scenarios. The first control-only scenario (CTR) assessed the control aspect for fixed design variables, meaning that the PV power and battery capacity are fixed. The second scenario, considering both the control and design (CTR & DES) aspects, allows for flexibility in designing the battery capacity and PV power. The two following subsections describe the results of the study performed in two steps, over the one-week and one-year data set, respectively.

## 3.1. A one-week toy example

In order to perform a simple comparative study, both CTR and CTR & DES analyses were conducted using data from a single summer week. This enables to optimize both methods on the same time horizon. This means training the RL algorithm on the same 168 time steps, with an initial day uniformly randomly selected over the week but an initial hour fixed at midnight. Additionally, during the training phase, the battery's initial SOC is uniformly sampled such that the RL algorithm is presented with a large variety of scenarios for improving the quality of the learned policy. The results for both the CTR scenario, where the design variables (i.e., the PV power and battery capacity) are fixed, and the CTR & DES scenario, where the PV and battery design variables are optimized in addition to control, are presented in Table 3.

| | | Unit | Optimal RL solution | Optimal MILP solution | MILP solution based on RL design |
|---|---|---|---|---|---|
| CTR | T | hours | 168 | 168 | |
| | Reward | € | -66 | -66 | |
| | Income | € | -30 | -30 | |
| CTR & DES | T | hours | 168 | 168 | 168 |
| | Reward | € | -40 | -46 | -53 |
| | Income | € | -4 | 0 | -17 |
| | Battery capacity | kWh | 62 | 40 | 62 |
| | PV power | kWp | 41 | 103 | 41 |

**Table 3**: Results of RL and MILP solutions on one-week data for control-only (CTR) and control and design (CTR & DES) scenarios. *T* denotes the time horizon in hours, while *income* represents the cost of buying and selling electricity from the grid, *reward* is the average cumulative reward value, and *battery capacity* and *PV power* indicate the values of design variables, which were set to 31.89 kWh and 55.81 kWp, respectively, in the CTR scenario. In the RL model, reward and income values were obtained by reloading the trained model with the determined design variables. Results were computed using an initial state of charge of the battery set to 50% of its capacity for both models. However, the RL model does not take into account the Eq. (10).

### 3.1.1. RL and MILP optimal objective values are similar in both scenarios but with different designs in the control and design scenario.

Table 3 shows that in the CTR scenario, the results of the RL approach are similar to those of MILP. This confirms that the DEPS algorithm is able to converge to the optimal solution of this specific problem. In the CTR & DES scenario, RL design variables differ from the MILP solution, resulting in an unexpected higher reward value (-40) than the MILP optimal one (-46). A detailed analysis reveals that this unexpected higher value is due to Eq. (10), which is not imposed in the MDP. In order to validate this analysis, the additional grid cost needed to fulfill Eq. (10) has been computed, taking into account the battery's final SOC obtained with the RL approach. As a result, the reward value has increased to -67 (instead of -40). This clearly highlights the importance that Eq. (10) plays in term of overall objective.

### 3.1.2. The CTR & DES scenario highlights differences in RL and MILP strategies.

It is seen from Table 3 that in the second scenario, the optimal design variables of the RL and MILP solutions differ. Finding different values in design variables shows that the DEPS algorithm is able to identify solutions with comparable reward but using different design strategies. In order to study the sensitivity of the optimal solution, the MILP method was applied by imposing the design variable values obtained with the RL, as it can be seen in the last column of Table 3. This indicates that the RL design solution is less optimal (-53) than the MILP one (-46).

## 3.2. A one-year case study

Optimal solutions of RL and MILP methods in both scenarios are now computed using data from a full year. The time horizon for the RL algorithm is still equal to $T = 168$, but the starting days are uniformly randomly selected over the year. The RL algorithm is trained over a pre-determined number of 100'000 iterations and the values of the RL design variables considered are the ones from the final iteration. The results are shown in Table 4.

### 3.2.1. The difficulty of generalizing a policy with stochasticity in the model and on the estimation of the expectation

It can be seen from Table 4 that in both the CTR and CTR & DES scenarios, the optimal reward obtained by the RL method is poorer than the MILP optimal rewards. Furthermore, as depicted in Fig. 2, due to the significant variations in the input data, the reward and income values exhibit substantial fluctuations across

| | | Unit | Optimal RL solution | Optimal MILP solution | MILP solution based on RL design |
|---|---|---|---|---|---|
| CTR | T | hours | 168 | 8760 | |
| | Reward | € | -268 | -228 | |
| | Income | € | -220 | -196 | |
| CTR & DES | T | hours | 168 | 8760 | 8760 |
| | Reward | € | -250 | -205 | -247 |
| | Income | € | -208 | -164 | -218 |
| | Battery capacity | kWh | 44 | 95 | 44 |
| | PV power | kWp | 57 | 81 | 57 |

**Table 4**: Results of RL and MILP solutions on one-year data for CTR and CTR & DES scenarios. $T$ denotes the time horizon in hours, while *income* represents the cost of buying and selling electricity from the grid, *reward* is the (expected) cumulative reward value, and *battery capacity* and *PV power* indicate the design variable values, which were set to 64.9 kWh and 63.65 kWp, respectively, in the CTR scenario. The RL solution is based on the trained model to determine the reward and income values, based on an average of 1'000 simulations over $T$ = 168. The MILP solution is computed for a one-year time horizon ($T$ = 8760). Both models use an initial state of charge (SOC) of the battery set to 50% of its capacity. However, the MILP model has an additional constraint specifies in Eq. (10).

iterations.

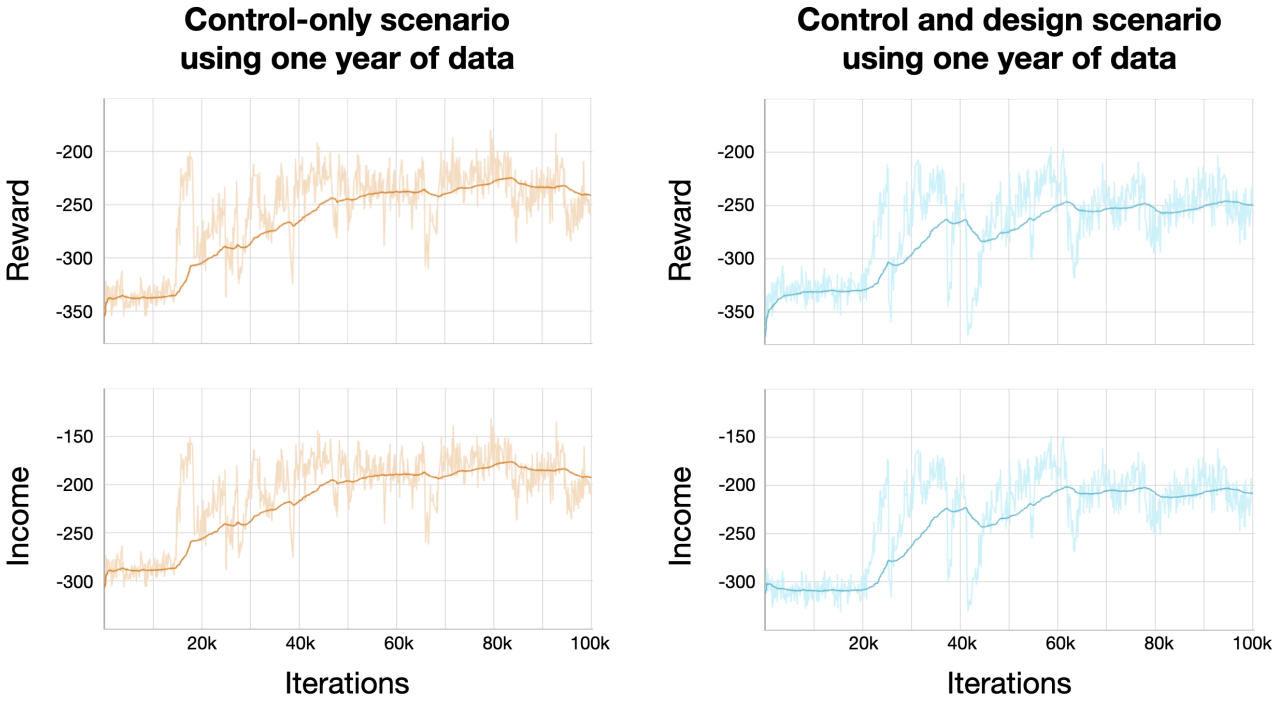


**Figure 2**: Value of reward and income obtained by the DEPS algorithm at each iteration for both scenarios. The left plots show the reward and income values for the CTR scenario, while the right plot shows the same values for the CTR & DES scenario. The light curve shows the exact values for each time step, while the dark curve displays the corresponding smoothed values. In the CTR scenario, the difference between the reward and income values remains constant at 31.93 due to fixed design variables, with a battery size of 64.9 kWh and a PV power of 63.65 kWp. However, in the CTR & DES scenario, the battery size and PV power output vary from 0 to 200 kWh and kWp, respectively.

During training in the CTR scenario (Fig. 2, left), the RL model achieved maximum reward and income values of -180 and -131, respectively, which are significantly better than the final results obtained from both methods in Table 4. This could suggest that depending on the set of weeks that are averaged at each iteration, it is possible to obtain a better or worse reward. Therefore, it seems important to work with a sufficiently representative number of weeks throughout the year. A similar observation can be made in the CTR & DES scenario, where the maximum reward and income values achieved were -195 and -148, respectively (Fig. 2, right).

#### 3.2.2. The RL method seems to promote lower design variable values

From Table 4 it is also seen that the RL approach seems to promote solutions involving lower values of design variables. To further investigate the reasons underlying this result, the design variables for the evolution of the battery capacity and PV power, during the training process, are reported in Fig. 3 in the CTR & DES scenario.

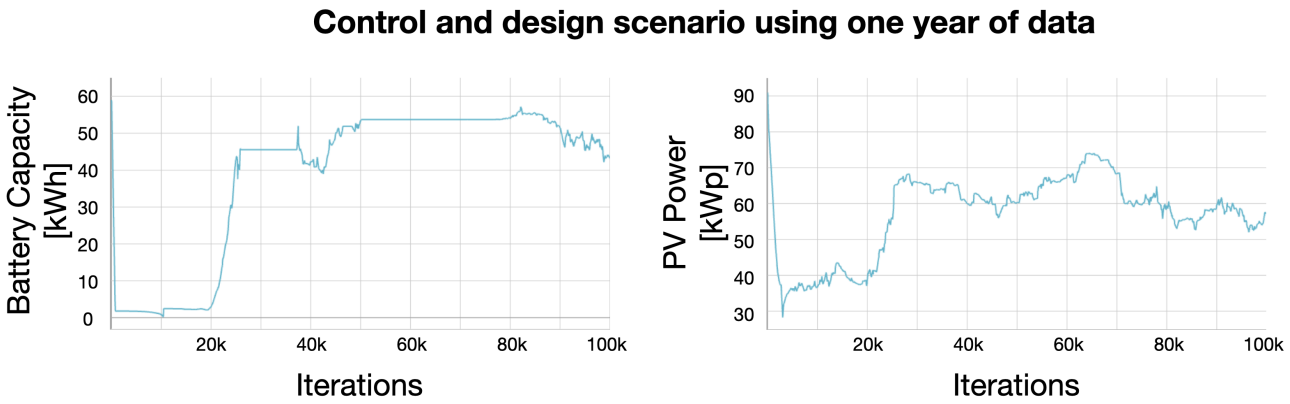


**Figure 3**: Value of design variables in the RL approach at each iteration, in the CTR & DES scenario. The RL algorithm converges at the final iteration to a battery capacity of approximately 44 kWh and a PV power output of around 57 kWp.

As indicated in Table 2, the design variable values can range from 0 to 200. However, it can be seen that higher values are not explored by the RL method. This latter resulted, at the last iteration, in design variables of 44 kWh for battery capacity and 57 kWp for PV power. During the training phase, the maximum values reached were 59 kWh and 90 kWp for battery capacity and PV power, respectively. This maximal explored battery capacity value is lower than the optimal one found by the MILP approach (95 kWh). Thus, the RL solution of the PV power value is expected to be lower. Indeed, the reward value is penalized if the RL agent injects PV production into the grid, since the cost of exported energy into the grid ($C_{grid}^{exp}$) is defined as a negative value in Table 2. Consequently, limited battery capacity intrinsically causes a lower PV power value.

## 4. Discussion

This section discusses the main observations that can be drawn from solving a battery-PV system with both RL and MILP approaches using the one-week and one-year data set.

### 4.1. The promises of RL for joint design and control of energy systems

The motivation for this study was to explore the potential of RL to enable joint control and design of energy systems. Tables 3 (one-week data) and 4 (one-year data) show that RL provides a solution that is close to the optimal MILP one. This is encouraging as it suggests that despite RL relying on a different optimization strategy, it is able to identify a meaningful solution in a simple case. However, the difference of reward value between MILP and RL increases when integrating design variables to the optimization problem, i.e., CTR & DES scenarios in Tables 3 and 4. Interestingly, the solutions for design variables are consistently smaller in RL as compared to MILP. Furthermore, from Fig. 3, it can be seen that the RL algorithm did not explore higher design variable values in the one-year case study. This observation can be explained by two possibilities: first, DEPS is a local-search method that is thus subject to converging towards local extrema. Once the control policy is too specialised to the investment parameters (under optimization too), these parameters are thus expected to be locally optimal and the algorithm is stuck. Second, the RL algorithms is subject to many hyperparameters to which the final results are sensible, it is possible that a different policy architecture, learning rate, or simply more iterations would ameliorate the performance of the method. Supporting the first explanation is the similarity between the reward values of the RL (-250) and MILP, based on same investments, (-247) approaches for the CTR & DES scenario with $T$ = 8760 (Table 4). Hence, in this specific energy system case study, it could be likely that the RL algorithm did not deem it advantageous to enhance the value of the design variables for either one or both of the two reasons stated.

Overall, these results show that RL provides realistic control and design strategies. Based on this, RL could be used to define new real-time control strategies integrating design constraints, and that are less sensitive to linearization inaccuracies [16], [17]. Given the differences in how uncertainties are accounted for by both methods, RL could also be a better candidate to integrate resources coming with high levels of uncertainty such as electric mobility.

### 4.2. Technical challenges and future directions

The main technical challenges encountered in this study are essentially the ones inherent to RL methods. First, various parameters need to be tuned: neural network architecture for the policy, the batch size for the optimization, the learning rate, or the different scaling among others. These parameters were tuned by trial and error and would need to be adapted to each new application. For example, the number of layers required in the one-year case study was larger than for the one-week toy example. Second, convergence of the RL method is not guaranteed, and when convergence happens, the solution is not guaranteed to be globally optimal. Third, as illustrated here above for the results of Figure 3, determining the number of iterations (set to 100,000 for the training phase in all our experiments) is also crucial and might affect RL solutions. Therefore, comparing RL and MILP solutions is not trivial because its is difficult to compare perfect foresight with policy based decisions. This should be accounted for when analyzing results from Tables 3 and 4.

From a technical point of view, future work will aim at using more advanced RL methods. In particular, the RL algorithm used here is a modified version of the REINFORCE algorithm [13], which was developed in 1992 and is one of the earliest RL algorithms. Today, more advanced algorithms are available for control problems, which can converge more rapidly or account for infinite time horizons, such as actor-critic algorithms (e.g., PPO [18] and GAE [19]), but are yet to be adapted to joint design and control. In terms of applications, future work will aim to better evaluate the added value of RL by assessing the long-term performance of real-time sized systems. For example, a control framework could be developed to establish an operation strategy for the MILP-sized system. The framework would then be evaluated using several years of real-time data from the same system used for design. The same exercise would be applied to the trained model of the DEPS algorithm and performance obtained from several years of system control would be benchmarked, and the impact of design decisions could be discussed with more perspective.

## 5. Conclusions

In most studies, MILP is used for the design of energy systems and RL for the control. On the one hand, MILP assumes a perfect foresight of the future and is difficult to generalize to new data. On the other hand, RL methods proved to be efficient in other tasks linked to design and control but not on energy systems. In this study, we assessed the potential of an RL method, DEPS, i.e. an RL algorithm proven efficient for designing and controlling complex systems, for the joint design and control of energy systems.

The energy system studied is a PV-battery system used to answer a real-life demand in order to minimize the overall cost. In order to assess the efficiency of the RL method, we compared the outcomes with those obtained with a MILP. As these two approaches are fundamentally different, the optimization problem was formulated in two distinct ways: first as a MILP and second as an MDP. The methodology and experimental context were clarified to facilitate the discussion of results and have a fair comparison. Both approaches are discussed in terms of their strengths and weaknesses.

The findings show that RL can produce control strategies that are close to optimal, while using different values of design variables. This highlights the potential of RL for joint design and control of energy systems, particularly in scenarios where stochasticity is a key factor. However, the study also highlights the difficulty of tuning and using theses methods. Moving forward, there are several challenges to address, including the need to ensure that the RL solution converges to a global optimum. However, the promising results obtained in this study suggest that RL has the potential to be a valuable tool for jointly designing and controlling energy systems.

# References


[1] A. Perera and P. Kamalaruban, "Applications of reinforcement learning in energy systems," *Renewable and Sustainable Energy Reviews*, vol. 137, p. 110 618, Mar. 2021. DOI: 10.1016/j.rser.2020.110618.

[2] A. T. D. Perera, P. U. Wickramasinghe, V. M. Nik, and J.-L. Scartezzini, "Introducing reinforcement learning to the energy system design process," en, *Applied Energy*, vol. 262, p. 114 580, Mar. 2020, ISSN: 0306-2619. DOI: 10.1016/j.apenergy.2020.114580. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0306261920300921 (visited on 10/24/2022).

[3] S. Fazlollahi and F. Maréchal, "Multi-objective, multi-period optimization of biomass conversion technologies using evolutionary algorithms and mixed integer linear programming (MILP)," en, *Applied Thermal Engineering*, Combined Special Issues: ECP 2011 and IMPRES 2010, vol. 50, no. 2, pp. 1504–1513, Feb. 2013, ISSN: 1359-4311. DOI: 10.1016/j.applthermaleng.2011.11.035. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S1359431111006636 (visited on 11/14/2022).

[4] A. Majid, S. Saaybi, T. Rietbergen, *et al.*, *Deep Reinforcement Learning Versus Evolution Strategies: A Comparative Survey*. May 2021. DOI: 10.36227/techrxiv.14679504.v1.

[5] H. Quest, M. Cauz, F. Heymann, *et al.*, "A 3D indicator for guiding AI applications in the energy sector," en, *Energy and AI*, vol. 9, p. 100 167, Aug. 2022, ISSN: 2666-5468. DOI: 10.1016/j.egyai.2022.100167. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2666546822000234 (visited on 11/03/2022).

[6] H. M. Abdullah, A. Gastli, and L. Ben-Brahim, "Reinforcement Learning Based EV Charging Management Systems–A Review," *IEEE Access*, vol. 9, pp. 41 506–41 531, 2021, Conference Name: IEEE Access, ISSN: 2169-3536. DOI: 10.1109/ACCESS.2021.3064354.

[7] M. Dorokhova, Y. Martinson, C. Ballif, and N. Wyrsch, "Deep reinforcement learning control of electric vehicle charging in the presence of photovoltaic generation," *Applied Energy*, vol. 301, p. 117 504, Nov. 2021. DOI: 10.1016/j.apenergy.2021.117504.

[8] W. Shi and V. W. Wong, "Real-time vehicle-to-grid control algorithm under price uncertainty," in *2011 IEEE International Conference on Smart Grid Communications (SmartGridComm)*, Oct. 2011, pp. 261–266. DOI: 10.1109/SmartGridComm.2011.6102330.

[9] W. Uther, "Markov Decision Processes," en, in *Encyclopedia of Machine Learning*, C. Sammut and G. I. Webb, Eds., Boston, MA: Springer US, 2010, pp. 642–646, ISBN: 978-0-387-30164-8. DOI: 10.1007/978-0-387-30164-8_512. [Online]. Available: https://doi.org/10.1007/978-0-387-30164-8_512 (visited on 03/10/2023).

[10] Manu Lahariya, N. Sadeghianpourhamami, and Chris Develder, "Computationally efficient joint coordination of multiple electric vehicle charging points using reinforcement learning," [Online]. Available: arXiv:2203.14078.

[11] N. Sadeghianpourhamami, J. Deleu, and C. Develder, "Definition and Evaluation of Model-Free Coordination of Electrical Vehicle Charging With Reinforcement Learning," *IEEE Transactions on Smart Grid*, vol. 11, no. 1, pp. 203–214, Jan. 2020, Conference Name: IEEE Transactions on Smart Grid, ISSN: 1949-3061. DOI: 10.1109/TSG.2019.2920320.

[12] A. Bolland, I. Boukas, M. Berger, and D. Ernst, "Jointly Learning Environments and Control Policies with Projected Stochastic Gradient Ascent," en, *Journal of Artificial Intelligence Research*, vol. 73, pp. 117–171, Jan. 2022, ISSN: 1076-9757. DOI: 10.1613/jair.1.13350. [Online]. Available: https://www.jair.org/index.php/jair/article/view/13350 (visited on 03/07/2023).

[13] R. J. Williams, "Simple statistical gradient-following algorithms for connectionist reinforcement learning," en, *Machine Learning*, vol. 8, no. 3, pp. 229–256, May 1992, ISSN: 1573-0565. DOI: 10.1007/BF00992696. [Online]. Available: https://doi.org/10.1007/BF00992696 (visited on 03/07/2023).

[14] B. Miftari, M. Berger, H. Djelassi, and D. Ernst, "GBOML: Graph-Based Optimization Modeling Language," en, *Journal of Open Source Software*, vol. 7, no. 72, p. 4158, Apr. 2022, ISSN: 2475-9066. DOI: 10.21105/joss.04158. [Online]. Available: https://joss.theoj.org/papers/10.21105/joss.04158 (visited on 03/07/2023).

[15] Gurobi, *Gurobi - The fastest solver*, Library Catalog: www.gurobi.com, 2020. [Online]. Available: https://www.gurobi.com/ (visited on 03/25/2020).

[16] M. Reuß, L. Welder, J. Thürauf, *et al.*, "Modeling hydrogen networks for future energy systems: A comparison of linear and nonlinear approaches," en, *International Journal of Hydrogen Energy*, vol. 44, no. 60, pp. 32 136–32 150, Dec. 2019, ISSN: 0360-3199. DOI: 10.1016/j.ijhydene.2019.10.080. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0360319919338625 (visited on 10/24/2022).

[17] C. Sánchez, L. Bloch, J. Holweger, C. Ballif, and N. Wyrsch, "Optimised Heat Pump Management for Increasing Photovoltaic Penetration into the Electricity Grid," *Energies*, vol. 12, p. 1571, Apr. 2019. DOI: 10.3390/en12081571.

[18] J. Schulman, F. Wolski, P. Dhariwal, A. Radford, and O. Klimov, *Proximal Policy Optimization Algorithms*, arXiv:1707.06347 [cs], Aug. 2017. [Online]. Available: http://arxiv.org/abs/1707.06347 (visited on 03/12/2023).

[19] J. Schulman, P. Moritz, S. Levine, M. Jordan, and P. Abbeel, *High-Dimensional Continuous Control Using Generalized Advantage Estimation*, arXiv:1506.02438 [cs], Oct. 2018. [Online]. Available: http://arxiv.org/abs/1506.02438 (visited on 03/12/2023).