

**Wydział Elektrotechniki, Automatyki, Informatyki i Inżynierii Biomedycznej**

Praca dyplomowa

*Zastosowanie uczenia ze wzmocnieniem w celu zastąpienia regulatora instalacji grzewczych*

*Reinforcement Learning Applications to Control of Heating Systems*

Autor: inż. *Adrian Gacek*

Kierunek studiów: Automatyka i Robotyka

Opiekun pracy: *dr inż. Jakub Grela*

Kraków, 2024

Spis treści

[1. Wstęp 3](#_Toc167624727)

[2. Podstawy teoretyczne 4](#_Toc167624728)

[2.1. Uczenie maszynowe 4](#_Toc167624729)

[Uczenie ze wzmocnieniem 4](#_Toc167624730)

[Agent i środowisko 5](#_Toc167624731)

[Podstawowe pojęcia w RL 5](#_Toc167624732)

[Proces decyzyjny Markova 6](#_Toc167624733)

[Funkcja wartości 7](#_Toc167624734)

[Eksploracja, a eksploatacja 8](#_Toc167624735)

[Model środowiska 8](#_Toc167624736)

[Off/on-policy/actor-critic 8](#_Toc167624737)

[I.i.e, batch learning 8](#_Toc167624738)

[3. Regulacja instalacji grzewczych 9](#_Toc167624739)

[3.1. Dotychczasowe rozwiązania 9](#_Toc167624740)

[3.2. Środowisko TCLab 11](#_Toc167624741)

[4. Algorytmy uczenia ze wzmocnieniem 11](#_Toc167624742)

[Q-learning 11](#_Toc167624743)

[Deterministic policy gradient 13](#_Toc167624744)

[Inne algorytmy 15](#_Toc167624745)

[5. Narzędzia implementacji 15](#_Toc167624746)

[Python 16](#_Toc167624747)

[Tensorflow 16](#_Toc167624748)

[Tf-Agents 16](#_Toc167624749)

[Optuna 16](#_Toc167624750)

[TCLab 16](#_Toc167624751)

[6. Bibliografia 17](#_Toc167624752)

# Wstęp

W ciągu ostatnich lat, można zaobserwować znaczące przyśpieszenie rozwoju sztucznej inteligencji, a jej potencjał znajduje zastosowanie w coraz szerszym zakresie problemów. Od generowania i identyfikacji obrazów, przez chatboty, po systemy sterowania, algorytmy uczenia maszynowego coraz częściej odgrywają kluczową rolę w rozwiązywaniu zarówno prostych, jak i niezwykle złożonych wyzwań. Jednym z obszarów, w którym sztuczna inteligencja staje się coraz bardziej znacząca, jest skuteczne i zoptymalizowane sterowanie systemami grzewczymi.

Współczesne systemy grzewcze odgrywają istotną rolę w zapewnieniu komfortu i bezpieczeństwa w budynkach, a ich odpowiednia regulacja jest kluczowa dla efektywnego wykorzystania energii. Ze względu na powszechność tego problemu oraz duże znaczenie ogrzewania w kontekście kosztów eksploatacji budynku, nawet niewielkie poprawy w sterowaniu mogą mieć istotny wpływ zarówno na budżet, jak i na środowisko.

W świetle powyższego, niniejsza praca ma na celu przeprowadzenie analizy możliwości wykorzystania algorytmów uczenia ze wzmocnieniem w problemach sterowania systemami grzewczymi. Użycie tych zaawansowanych technik może prowadzić do bardziej efektywnego wykorzystania energii, poprawy komfortu użytkowników oraz redukcji negatywnego wpływu na środowisko naturalne.

Kolejne rozdziały będą miały na celu kolejno przedstawienie teoretycznych podstaw zagadnienia uczenia maszynowego, w szczególności przegląd metod uczenia ze wzmocnieniem i istotnych pojęć z nimi związanych, oraz zaprezentowane zostaną przykłady zastosowania podobnych rozwiązań w literaturze. Następnie w części praktycznej przedstawiony zostanie proces przygotowania środowiska opartego o symulację TCLab i uczenia wskazanych w dalszych rozdziałach algorytmów. Na koniec nakreślono wyniki testów symulacyjnych oraz sprzętowych, wraz z wnioskami.

# Podstawy teoretyczne

Niniejszy rozdział ma za zadanie przedstawić podstawowe idee stojące za uczeniem maszynowym i metodami uczenia ze wzmocnieniem.

## Uczenie maszynowe

Uczenie maszynowe to obszerna dziedzina sztucznej inteligencji, zajmująca się rozwijaniem algorytmów i technik umożliwiających komputerom uczenie się z danych i doświadczenia, bez konieczności jawnego programowania. Polega ono na tworzeniu modeli, które analizując dane, wyszukują wzorców i podejmują decyzje lub prognozy na ich podstawie.

Uczenie maszynowe można podzielić na następujące obszary:

* Uczenie nadzorowane
* Uczenie nienadzorowane
* Uczenie ze wzmocnieniem

Algorytmy uczenia nadzorowanego otrzymują zestaw danych wejściowych, w którym każdy przykład posiada etykietę określającą pożądane wyjście. Celem jest znalezienie funkcji mapującej dane wejściowe na dane wyjściowe. Są one stosowane na przykład w problemach klasyfikacji i regresji.

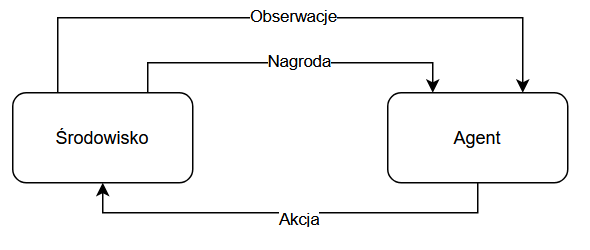
Uczenie nienadzorowane zakłada, że algorytm otrzymuje dane wejściowe bez etykiet i ma za zadanie znajdować wzorce występujące w danych. Taki rodzaj algorytmów może być zastosowany do klasteryzacji, czy wykrywania anomalii.

## Uczenie ze wzmocnieniem

Uczenie ze wzmocnieniem (w skrócie RL, z angielskiego Reinforcement Learning) to proces uczenia się, w którym celem jest maksymalizacja nagrody poprzez przyporządkowanie sytuacji do akcji. Uczący się algorytm nie otrzymuje wskazówek, które akcje wykonać, lecz nabiera doświadczenia poprzez przeprowadzenie interakcji z systemem. Tworzy się w ten sposób model wskazujący, które akcje przynoszą największą nagrodę, często biorąc pod uwagę nie tylko stan aktualny, ale również następujące po nim stany przyszłe. Poszukiwanie metod prób i błędów oraz opóźniona nagroda to dwie główne cechy odróżniające uczenie ze wzmocnieniem od innych form uczenia. Problematyka uczenia ze wzmocnieniem jest formalizowana przy użyciu teorii systemów dynamicznych. [1]

## Agent i środowisko

Proces uczenia ze wzmocnieniem opiera się o interakcję między agentem a środowiskiem: agent wykonuje akcję, na którą reaguje środowisko, zmieniając swój stan i przyznając za to nagrodę. Następnie agent otrzymuje nagrodę oraz obserwacje zawierające informacje o stanie środowiska, które wykorzystuje do dalszej nauki i optymalizacji swoich przyszłych decyzji. Celem jest opracowanie strategii, która maksymalizuje sumę przyszłych nagród, które agent może otrzymać.



Rys. 1. Interakcja między środowiskiem i agentem. Opracowanie własne przy pomocy narzędzi draw.io

## Podstawowe pojęcia w RL

Dla pełnego zrozumienia procesu uczenia ze wzmocnieniem należy zdefiniować jego podstawowe elementy [1]:

**Obserwacje** – są to wszystkie informacje o stanie środowiska, które umożliwiają określenie następnej akcji podjętej przez agenta. Warto zauważyć, że nie w każdym systemie agent ma dostęp do wszystkich zmiennych stanu, co może utrudnić proces uczenia.

**Akcja** – może być zdefiniowana w sposób zarówno ciągły, jak i dyskretny. Umożliwia wpływanie agenta na zmianę stanu środowiska.

**Nagroda** - zmienna zwracana przez środowisko po wykonaniu kroku przez agenta, obrazuje to, jak korzystne było wykonanie akcji. Celem agenta jest maksymalizowanie sumarycznej nagrody gromadzonej podczas całego cyklu życia procesu.

**Trajektoria –** ciąg następujących po sobie trójek stan-akcja-nagroda reprezentujący interakcję agenta ze środowiskiem: ()

**Polityka –** funkcja mapująca obserwacje ze środowiska na akcje. Ma za zadanie zdefiniować zachowanie agenta w sytuacjach, w jakich może znaleźć się środowisko. Znalezienie odpowiedniej polityki jest celem uczenia ze wzmocnieniem. Może być definiowana zarówno deterministycznie, jak i stochastycznie.

**Stopę dyskontową** – współczynnik określający relację między natychmiastową nagrodą, a przewidywaną sumą przyszłych nagród. Wzór (1.1) wskazuje typowe zastosowanie:

|  |  |  |
| --- | --- | --- |
|  |  | (1.1) |

Gdzie:

* – maksymalizowana w procesie uczenia wartość sumy nagród oceniająca korzyść z wykonania danej akcji, nazywana również zwrotem
* numer stanu, dla którego wyznaczamy wartość sumy nagród
* – stopa dyskontowa, wartość pomiędzy zapewnia, że każdy następny przewidywany stan, ma mniejsze znaczenie względem natychmiastowej nagrody
* – numer przyszłego przewidywanego stanu
* – Przyszła przewidywana nagroda

## Proces decyzyjny Markova

U podstaw paradygmatu uczenia ze wzmocnieniem leży proces decyzyjny Markova (Markov Decision Process, MDP), z pomocą którego można formalnie zdefiniować wiele badanych w tej dziedzinie problemów. Mając na uwadze elementy wymienione w poprzednim podrozdziale warto omówić cechy takiego procesu.

Proces decyzyjny Markova jest zdefiniowany jako piątka (𝑆, 𝐴, 𝒯, 𝑟, 𝛾) [16].

Gdzie:

* 𝐴 – przestrzeń akcji.
* 𝑆 – przestrzeń stanów
* 𝒯 – dynamika przejść - model systemu, który opisuje zasady rządzące przejściami z jednego stanu do drugiego. (W stanie 𝑠𝑡 podejmowana jest akcja 𝑎𝑡,   
  która zmienia system na nowy stan 𝑠𝑡+1∈𝑆, zgodnie z dynamiką przejść   
  środowiska 𝒯(𝑠𝑡+1|𝑠, 𝑎).
* 𝑟 – funkcja nagrody 𝑟: 𝑆×𝐴⟶ℝ
* 𝛾 – stopa dyskontowa

Bardzo istotną cechą takiego procesu jest to, że spełnia on własność Markova. Pomimo, że nagrody, jak i przejścia mogą być określone przez prawdopodobieństwo, zależą one tylko i wyłącznie od aktualnego stanu i akcji. Poprzednie wartości nie mają wpływu na przyszłość. Warto jednak nadmienić, że jest to cecha modelu opisującego proces. Czy dany proces będzie spełniał taki warunek zależy od sposobu w jaki została zdefiniowana przestrzeń stanu. [18]

Celem agenta jest maksymalizacja oczekiwanej zdyskontowanej nagrody poprzez wybieranie odpowiednich akcji. Agent powinien nauczyć się optymalnej polityki jak wskazano we   
wzorze (1.2) [16] .

|  |  |  |
| --- | --- | --- |
|  |  | (1.2) |

Gdzie *Ea~π*  oznacza oczekiwaną wartość zwrotu, w wyniku wyboru akcji 𝑎, na podstawie polityki *π* .

## Funkcja wartości

Funkcja wartości określa, co jest korzystne w dłuższej perspektywie czasowej, w przeciwieństwie do sygnału nagrody, który ocenia tylko natychmiastowe efekty akcji. W przybliżeniu wartość stanu to całkowita suma nagród, jakiej agent może oczekiwać, zaczynając od danego stanu. Ma na celu wskazać na długoterminową atrakcyjność stanów, biorąc pod uwagę stany, które prawdopodobnie nastąpią oraz nagrody dostępne w tych stanach. Wzór (1.3), przedstawia jak funkcja wartości określa oczekiwany zwrot przy zastosowaniu   
polityki π. [1]

|  |  |  |
| --- | --- | --- |
|  |  | (1.3) |

Podobnie określa się również funkcję wartości akcji [wzór (1.4)], czyli funkcja wartości przy założeniu, że w aktualnym stanie zostanie wykonana akcja : [1]

|  |  |  |
| --- | --- | --- |
|  |  | (1.4) |

## Eksploracja, a eksploatacja

Istotnym problemem w dziedzinie RL jest problem balansu między eksploracją i eksploatacją. Wynika on z faktu, że zbieranie informacji i ich wykorzystywanie to często dwa wzajemnie wykluczające się działania. Z jednej strony, zależy nam, aby agent eksplorował, czyli zdobywał informacje o swoim otoczeniu w celu poprawy przyszłych wyników. Ponosi jednak w ten sposób koszty zdobywania nowych informacji nie korzystając z posiadanej wiedzy. Z drugiej strony, eksploatując czyli używając wyłącznie istniejącej wiedzy, ograniczamy zdolności adaptacji do istotnych zmian w otoczeniu i system może utknąć w suboptymalnej stabilnej równowadze. Dlatego optymalne zachowanie zwykle wymaga pewnej równowagi między działaniami eksploracyjnymi a eksploatacyjnymi. [2]

Przykładem rozwiązania tego problemu może być jedna z prostych metod balansowania między eksploracją a eksploatacją ε-greedy. Działa ona w następujący sposób [3]:

* **Eksploatacja**: Większość czasu metoda ε-greedy działa zachłannie (greedy), czyli wybiera działanie, które w danej chwili wydaje się najlepsze na podstawie dostępnej wiedzy.
* **Eksploracja**: Co pewien czas, z małym prawdopodobieństwem ε, metoda wybiera losowo jedno z dostępnych działań, zamiast wybierać działanie uważane za najlepsze.

## Model środowiska

Jednym z popularnych sposobów podziału metod uczenia ze wzmocnieniem jest ten, oparty o wiedzę agenta na temat modelu środowiska.

**Pierwszą** kategorią są metody **model-based**, czyli oparte na modelu środowiska. Model ten umożliwia przewidywanie przyszłych stanów i nagród na podstawie aktualnego stanu i podjętych akcji, dzięki symulacji zachowania środowiska. W ten sposób, można planować działania, rozważając możliwe przyszłe sytuacje przed ich faktycznym doświadczeniem. **Metody Model-free**, są pozbawione modelu środowiska. Uczą się na zasadzie prób i błędów, bez posiadania modelu środowiska. Są one prostsze i nie wykorzystują planowania, zamiast tego uczą się bezpośrednio na podstawie interakcji ze środowiskiem. Nowoczesne systemy RL mogą łączyć oba podejścia. [1]

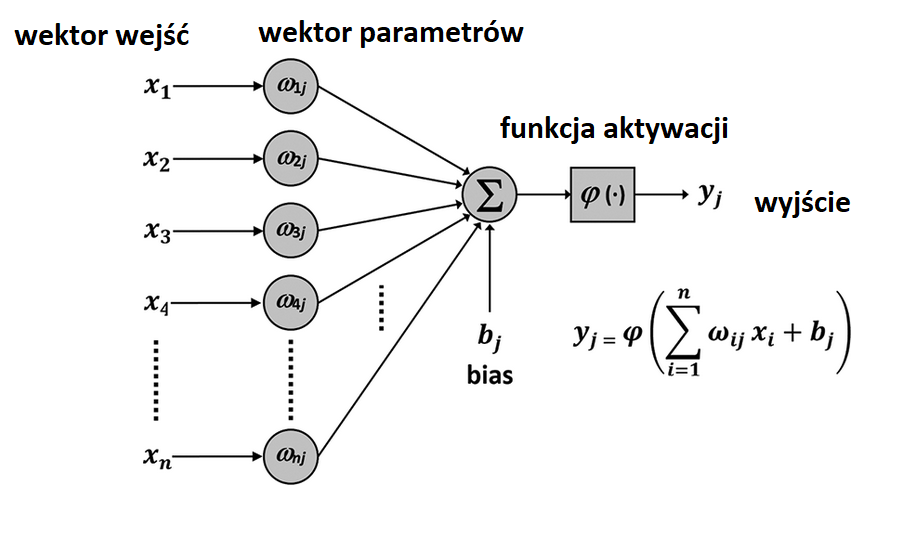
## Off-policy i on-policy

Agenci w procesie treningu mogą być uczeni na dwa sposoby. W przypadku metod   
**on-policy**, algorytm posiada jedną politykę, która jest jednocześnie optymalizowana, oraz wykorzystywana do interakcji z otoczeniem. Metody **off-policy** posiadają osobną politykę do generowania danych ze środowiska i osobną tzw. docelową (target policy), która jest uczona. Próbki generowane przez tzw. behavioral policy są przechowywane w buforze doświadczeń (ang. Experience Replay) podczas interakcji ze środowiskiem. Podczas treningu próbki z bufora są używane do aktualizacji polityki docelowej. Dane treningowe mogą być wygenerowane na podstawie starszej wersji polityki, dlatego są one rozpatrywane osobno. [22]

W literaturze można również znaleźć przykłady połączenia tych dwóch metod. Grupa takich rozwiązań nazywana jest aktor-krytyk (actor-critic), ponieważ w procesie treningu uczona jest zarówno polityka, jak i funkcja wartości, która jest używana do jej oceny. [22]

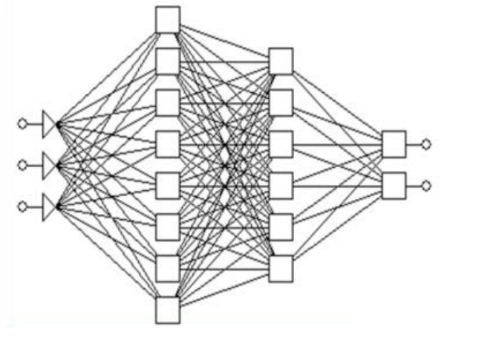
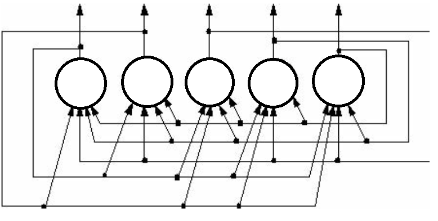
## Sieci neuronowe

Sieci neuronowe są systemem inspirowanym działaniem ludzkiego mózgu podczas przetwarzaniu informacji. Są one zbudowane z połączonych sztucznych neuronów, które przetwarzają dane wejściowe, przekazując je przez serię warstw, aby wygenerować pożądane wyjście. Podstawowymi jednostkami sieci neuronowych są sztuczne neurony. Rys. 2. przedstawia najprostszą sieć składającą się z jednego neuronu. Każdy neuron przyjmuje wiele wejść, przetwarza je za pomocą funkcji aktywacji i generuje sygnał, który jest przekazywany do innych neuronów lub do wyjścia. Każde połączenie z wejściem ma przypisaną wagę, która modyfikuje wpływ sygnału. [31] Wagi są parametrami, które sieć neuronowa uczy się dostosowywać podczas procesu uczenia. Funkcje aktywacji (np. softmax, ReLU, sigmoida) decydują, czy i w jakim stopniu neuron zostanie aktywowany.



Rys. 2. Schemat perceptronu [30]

W większości problemów sieci składają się z wielu neuronów, najczęściej połączonych w warstwy (Rys. 3). W takich przypadkach w procesie optymalizacji parametrów stosowany jest algorytm propagacji wstecznej, aby umożliwić dostosowywanie wag w warstwach oddalonych od wyjścia, dla którego wyznaczany jest błąd względem oczekiwanych rezultatów. Propagacja wsteczna (backpropagation) działa na zasadzie propagacji błędu od warstwy wyjściowej wstecz do warstw ukrytych, przy czym obliczane są pochodne cząstkowe funkcji kosztu względem wag na podstawie reguły łańcuchowej.



Rys. 3. Przykłady różnych struktur połączeń neuronów. Sieci Hopfielda (po lewej), połączenia każdy z każdym   
(po prawej). [31]

Na przestrzeni lat powstało wiele rodzajów sieci, różniących się strukturą i zastosowaniem. Najprostszym przykładem mogą być wielowarstwowe perceptrony, w których sygnał porusza się w tylko jednym kierunku i każdy neuron kolejnej warstwy jest połączony ze wszystkimi neuronami warstwy poprzedniej. Sieci konwolucyjne składają się między innymi z warstw konwolucyjnych, które symulują działanie filtra przesuwnego, dzięki połączeniu neuronu z fragmentem poprzedniej warstwy należącym do pewnego otoczenia. Takie sieci dobrze sprawdzają się w analizie problemów, w których wejścia mają strukturę przestrzenną, jak obrazy czy szeregi czasowe. Sieci rekurencyjne wykorzystują dodatkowe połączenia wewnątrz warstw, tworząc pętle, które przechowują informacje o poprzednich stanach w jakich znajdowały się neurony. Szczególnie użyteczne są w analizie sekwencji. W RL różne sieci neuronowe znalazły zastosowanie między innymi przy aproksymacji funkcji wartości [13] (sieć rekurencyjna GRU), polityki [24] (wielowarstwowy perceptron). czy modeli środowiska [29] (wielowarstwowy perceptron).

## Bufor doświadczeń

Większość algorytmów optymalizacyjnych zakłada, że próbki są niezależne i identycznie rozłożone. Pobieranie doświadczeń w sposób sekwencyjny, powoduje jednak, że sąsiednie stany są ze sobą ściśle skorelowane. Dodatkowo, aby efektywnie wykorzystać optymalizację sprzętową, istotne jest stosowanie batchy, a więc zestawów doświadczeń, zamiast pobierania ich online sekwencyjnie. W celu rozwiązania tego problemu stosuje się bufor powtórek. Jest to skończony bufor o ograniczonym rozmiarze przechowujący krotki (). Pobieranie zestawów np. w sposób losowy zapewnia niezależność próbek wymaganą między innymi w optymalizacji sieci neuronowych. [24]

# Regulacja instalacji grzewczych

Instalacje grzewcze odgrywają kluczową rolę w zapewnieniu komfortu cieplnego i efektywności energetycznej budynków. W dobie rosnącej świadomości ekologicznej oraz wzrastających kosztów energii, optymalna regulacja tych systemów staje się nie tylko pożądana, ale wręcz niezbędna. Odpowiednia regulacja takiej instalacji polega na dostosowywaniu parametrów pracy urządzeń grzewczych w celu osiągnięcia zamierzonych warunków cieplnych przy minimalnym zużyciu energii. Ponieważ jest to problem bardzo szeroko rozpowszechniony, potencjalne korzyści w zakresie poprawy jakości życia, ograniczenia kosztów, oraz zmniejszenia niekorzystnego wpływu człowieka na środowisko mogą być znaczące.

Niniejszy rozdział przedstawi przykładowe rozwiązania proponowane przez badaczy związanych z dziedziną regulacji ogrzewania, zwłaszcza w obszarze ciepłej wody użytkowej. Dodatkowo zostanie w nim zaprezentowane środowisko, które zostanie wykorzystane do praktycznych testów algorytmów.

## 3.1. Dotychczasowe rozwiązania

Bardzo popularnymi metodami sterowania instalacjami grzewczymi są systemy wykorzystujące regulatory PID. Typowy regulator PID działa w pętli sprzężenia zwrotnego, poprzez trzy podstawowe mechanizmy. Część proporcjonalna (P) generuje sygnał proporcjonalny do bieżącego błędu sterowania. Część całkująca (I) uwzględnia sumę wszystkich wcześniejszych błędów w czasie, co eliminuje stały błąd. Część różniczkująca (D) reaguje na szybkość zmiany błędu, pomagając w tłumieniu oscylacji i poprawie stabilności systemu. Jest to bardzo proste i wydajne rozwiązanie, jednak w wielu przypadkach, wyniki mogą być niezadowalające. W nowych publikacjach, regulator PID często jest łączony z innymi metodami, aby osiągać lepsze wyniki. Jednym z przykładów może być praca [6], w której autorzy prezentują porównanie 3 kontrolerów w systemie elektrycznego zbiornika wody grzewczej. Dzięki zastosowaniu logiki rozmytej możliwa była znacząca poprawa wyników, zarówno pod względem czasu, jak i dokładności regulacji. Podobnych rozwiązań można znaleźć wiele więcej: [5], [6], [7].

Kontrolery oparte o logikę rozmytą (z ang.  *fuzzy logic controller*), również są bardzo popularną metodą sterowania, co pokazała już część prac wymienionych powyżej. Takie regulatory wykorzystują logikę wielowartościową. Sygnały wejściowe takiego regulatora poddawane są w pierwszej kolejności fuzzifikacji, czyli określany jest ich stopień przynależności do zbiorów rozmytych sygnału wejściowego, by następnie stopień ten zmapować, na podstawie bazy reguł na stopień przynależności do zbiorów rozmytych sygnału wyjściowego. Na koniec dokonywana jest defuzzifikacja, otrzymanego sygnału, co jest operacją odwrotną do fuzzifikacji. [8] Zastosowanie takiego mechanizmu można wykorzystać zarówno samodzielnie ([4], [8]), jak i we współpracy z innymi metodami (z PID: [6], z uczeniem maszynowym: [9]). Do głównych zalet kontrolerów logiki rozmytej należy zdolność radzenia sobie z nieprecyzyjnymi i nieliniowymi systemami, oferując większą elastyczność i bardziej intuicyjne projektowanie dzięki użyciu reguł opartych na języku naturalnym. Jednakże, ich wady obejmują trudności w doborze odpowiednich funkcji przynależności i reguł oraz potencjalnie większe wymagania obliczeniowe w porównaniu do klasycznych regulatorów PID.

Powyższe rozwiązania mogą jednak być niewystarczające w przypadku bardziej złożonych systemów oraz gdy zależy nam na większych oszczędnościach energii. W takich sytuacjach przydatne mogą okazać się bardziej zaawansowane propozycje. Na przykład przedstawiony w pracy [10] algorytm pozwala na zaoszczędzenie od 8% do 28% energii, dzięki zastosowaniu algorytmu przewidującego obecność osób w pomieszczeniach. Autorzy porównywali wyniki z tradycyjnym systemem regulacji temperatury opartej o timer.

Thilker i in. w artykule [11] wprowadzili koncepcję predykcyjnego sterowania nieliniowego modelu na potrzeby inteligentnego ogrzewania budynków, w którym wykorzystuje się inteligentne termostaty i czujników internetu rzeczy w celu umożliwienia precyzyjnego sterowania systemami grzewczymi za pomocą sterowania predykcyjnego. Podejście to oznacza przejście w stronę bardziej wyrafinowanych technik sterowania w celu poprawy wydajności i szybkości reakcji systemów grzewczych.

W systemach ogrzewania, obiecującym podejściem do zwiększania efektywności energetycznej, optymalizacji wydajności i zapewniania komfortu użytkowników okazało się wykorzystanie uczenia maszynowego. W proponowanej w pracy [12] metodologii zastosowano model uczenia maszynowego, który wykorzystuje zarówno zmierzone dane (np. temperatura wewnętrzna/zewnętrzna, wilgotność względna), jak i dane prognozowane (np. dane meteorologiczne), aby trenować model regresji liniowej wielokrotnej do prognozowania temperatury w analizowanym pomieszczeniu. Następnie, przy użyciu metody optymalizacji algorytmem genetycznym, model ten ocenia różne strategie ogrzewania. Każdej strategii przypisuje się ocenę na podstawie zdefiniowanych przez użytkownika kryteriów, co pozwala na priorytetyzację i wybór najlepszej strategii.

W celu poprawy wydajności systemów HVAC (ang. heating, ventilation, air conditioning), w badaniu [13] zaproponowano metodę hybrydową WDQN-temPER, która łączy sieć deep Q-learning (DQN) z techniką priorytetowego odtwarzania doświadczeń (PER) oraz modelem jednostki rekurencyjnej (GRU). Model GRU przewiduje przyszłe temperatury zewnętrzne, które są wykorzystywane jako zmienne stanu w modelu RL (uczenie przez wzmacnianie), natomiast PER pozwala na efektywniejsze wykorzystywanie doświadczeń.

W artykule [14] autorzy proponują sformułowanie projektowania układu sterowania jako procesu decyzyjnego Markowa z użyciem głębokich sieci neuronowych i wykorzystanie algorytm głębokiego uczenia przez wzmacnianie opartego na głębokich gradientach polityki deterministycznej (DDPG) do znalezienia optymalnej strategii sterowania systemami HVAC, która równoważy koszty energii elektrycznej i komfort użytkowników.

Warto również wskazać pracę [15], w której Blad i in. zaproponowali uczenie ze wzmocnieniem wykorzystujące wielu agentów, w systemie HVAC, aby osiągnąć minimalny czas trenowania, działając on-line. Zaprezentowany algorytm został zaprojektowany, aby posiadał zdecentralizowaną strukturę, w której tylko istotne stany są udostępniane między agentami, a akcje są sekwencyjnie koordynowane. Większość agentów dobiera akcje na podstawie stanu środowiska, a na koniec tzw. mixing agent wybiera akcje, na podstawie stanu środowiska i akcji z pozostałych agentów.

## 3.2. Środowisko TCLab

# Algorytmy uczenia ze wzmocnieniem

Niniejszy rozdział ma za zadanie przedstawić przegląd algorytmów uczenia ze wzmocnieniem. Zaprezentowane zostaną ich ogólne zasady działania oraz najważniejsze cechy.

## Q-learning

Q-learning jest algorytmem zaproponowanym przez Chrisa Witkinsa w 1989 roku [18]. Jest to metoda typu model-free, off-policy. Jest podstawą na której opracowano wiele późniejszych rozwiązań, np. omówione niżej TD3 i DDPG. Opiera się na funkcji wartości akcji , która ocenia jakość wykonania danej akcji w stanie . Algorytm aktualizuje wartości tej funkcji na podstawie doświadczeń zdobywanych podczas interakcji z otoczeniem. Działanie algorytmu można przedstawić następująco [19]:

W każdym kroku uczenia *t* agent:

1. Obserwuje aktualny stan *t*
2. Wybiera i wykonuje akcję *at*
3. Obserwuje następny stan
4. Otrzymuje natychmiastowy zwrot nagrody *rt*
5. Dostosowuje funkcję według wzoru (3.1), *αt* – współczynnik uczenia (learning rate)

|  |  |  |
| --- | --- | --- |
|  |  | (3.1) |
|  | Gdzie: |  |
|  |  | (3.2) |

Z początku funkcja była reprezentowana przy pomocy tablicy look-up table, jednak dziś coraz popularniejsze staje się wykorzystywanie jej aproksymacji, co pozwala rozszerzyć obszar zastosowań do bardziej złożonych problemów, w tym także na problemy z ciągłą przestrzenią stanów. Szczególnie interesujące okazało się stosowanie do aproksymacji głębokich sieci neuronowych. Takie algorytmy nazywane są Deep Q-Learning (DQN). Problemem takiego rozwiązania jest jednak, to, że nie ma pewności co do zbieżności polityki. Warunkiem zbieżności algorytmu Q-learning jest fakt, że może przejść przez każdą parę stan-akcja nieskończoną ilość razy i learning rate jest odpowiednio redukowany [20]. Warunek ten jest spełniony tylko w przypadku dyskretnej przestrzeni stanów.

REINFORCE

REINFORCE to klasa algorytmów, która dokonuje optymalizacji polityki bezpośrednio poprzez maksymalizację oczekiwanej nagrody. Algorytm ten działa w ramach podejścia on-policy, a więc uczy się na podstawie polityki, którą aktualnie stosuje agent. Algorytmy REINFORCE dostosowują wagi w kierunku leżącym wzdłuż gradientu oczekiwanego wzmocnienia zarówno w zadaniach ze wzmocnieniem natychmiastowym, jak i w pewnych ograniczonych formach zadań ze wzmocnieniem opóźnionym, i robią to bez jawnego obliczania szacunków gradientu lub nawet przechowywania informacji, na podstawie których można by dokonać takich szacunków. [21] Metoda nie wymaga znajomości modelu środowiska przez agenta.

REINFORCE wykorzystuje do aktualizacji polityki oszacowane skumulowane zwroty uzyskane z zestawu próbkowanych trajektorii za pomocą metody Monte Carlo. Najczęściej używanym wariantem tego algorytmu jest forma z bazą, której celem jest zmniejszenie wariancji generowanej przy oszacowywaniu gradientu. Wzór (3.3) przedstawia wyznaczany w procesie uczenia gradient oczekiwanego zwrotu, który pozwala na poprawianie wag   
polityki. [22]

|  |  |  |
| --- | --- | --- |
|  |  | (3.3) |

Gdzie jest funkcją wartości niezależną od akcji .

Do zalet tej metody należą przede wszystkim względna prostota, oraz możliwość działania w ciągłej przestrzeni akcji. Istotną wadą może być jednak ryzyko zbiegania algorytmu do optimum lokalnego.

## Deterministic policy gradient

Podobnie jak REINFORCE algorytmy klasy Deterministic Policy Gradient (DPO) są metodami opartymi o gradienty polityki (policy gradient). Istotną różnicą jest jednak fakt, że opisane powyżej metody są stochastyczne w przeciwieństwie do deterministycznych przedstawionych przez Silvera i in. w pracy [23]. Algorytmy DPO są pozbawione modelu środowiska, jednak zaliczają się do nich zarówno rozwiązania off-policy, jak i on-policy. Zostały opracowane, aby rozwiązywać problemy ciągłej przestrzeni akcji.

W przypadku metod stochastycznych polityka mapuje stany na prawdopodobieństwa wykonania danej akcji. Metody deterministyczne są szczególnym przypadkiem takich metod, gdy, prawdopodobieństwo jednej z nich jest zawsze równe 1. Pozwala to na przyspieszenie wyznaczania , ponieważ nie trzeba sprawdzać wyników każdej akcji w następujących stanach. Gradient oczekiwanego zwrotu w tych metodach zależy przede wszystkim od gradientu polityki oraz gradientu funkcji wartości jak określono we wzorze (3.4) [23]:

|  |  |  |
| --- | --- | --- |
|  |  | (3.4) |

Gdzie:

– wektor parametrów deterministycznej polityki : S

- rozkład prawdopodobieństwa stanów dla polityki

Rozwinięciem DPG jest zaprezentowana w [24] Deep DPG (DDPG) metoda, która wykorzystuje głębokie sieci neuronowe do skalowania rozwiązania na bardziej złożone problemy. DDPG wprowadza wykorzystanie sieci docelowych zarówno dla aktora, jak i krytyka w celu stabilizacji treningu. Sieci docelowe są powoli aktualizowane, aby zapobiec gwałtownym wahaniom w procesie uczenia się. Rozwiązanie przedstawione w [24] proponuje również wykorzystanie strategii eksploracji, poprzez dodanie szumu Ornsteina-Uhlenbecka do polityki deterministycznej, aby zapewnić wystarczającą eksplorację przestrzeni działania, pozwala to otrzymać gładszy sygnał szumu, względem zupełnie losowych wartości, co umożliwia bardziej realistyczną eksplorację np. w środowiskach o pewnej bezwładności. Algorytmy DPG charakteryzują się istotnie wyższą wydajnością względem poprzednio omawianych rozwiązań.

Twin Delayed Deep Deterministic Policy Gradient

Twin Delayed Deep Deterministic Policy Gradient (TD3) to algorytm uczenia ze wzmocnieniem, który jest szczególnym przypadkiem metody Actor-Critic. TD3 został opracowany w celu poprawienia stabilności i wydajności algorytmu DDPG, zwłaszcza pod kątem przeceniania wartości z powodu błędów aproksymacji. Istotną cechą algorytmu jest zastosowanie dwóch sieci krytyków, w ten sposób do aktualizacji parametrów funkcji aproksymujących może zostać wykorzystana minimalna wartość Q. Drugim ważnym elementem jest opóźnienie aktualizacji polityki. Algorytm rzadziej dokonuje jej aktualizacji względem zmian w parametrach krytyków. Ponadto do akcji otrzymywanych z docelowej polityki dodawany jest szum w trakcie trenowania, ograniczając eksploatację błędów. Zastosowanie takich rozwiązań ogranicza przeszacowania wartości Q.

Algorytm TD3 pozwala na rozwiązywanie problemów z ciągłą przestrzenią akcji bez znajomości modelu środowiska.

Proximal policy optimization

Proximal Policy Optimization (PPO) [25] jest jedną z nowoczesnych metod uczenia ze wzmocnieniem. Oparta jest na gradientach polityki, łącząc zalety metod policy gradient i Trust Region Policy Optimization (TRPO), jednocześnie upraszczając implementację i zwiększając stabilność treningu. Jest to metoda off-policy, niewykorzystująca modelu środowiska.

PPO wprowadza funkcję celu, która ogranicza zmiany w polityce. W przeciwieństwie do tradycyjnych metod, które mogą powodować zbyt gwałtowne zmiany w polityce, PPO kontroluje te zmiany, co prowadzi do bardziej stabilnego uczenia. Proces nauki polega na poszukiwaniu wektora wag polityki, dla której oczekiwana wartość funkcji celu będzie największa. Wzór na funkcję celu (3.5) [25]:

|  |  |  |
| --- | --- | --- |
|  |  | (3.5) |

Gdzie:

* - PPO używa operacji "clipping" w celu ograniczenia stosunku prawdopodobieństw nowej polityki względem starszej. Operacja ta ogranicza wartości stosunku do zakresu [1−,1+], co zapobiega zbyt dużym zmianom w polityce.
* – tzw. Advantage Function (3.6):

|  |  |
| --- | --- |
|  | (3.6) |

PPO jest stabilniejszy niż tradycyjne metody gradientów polityki dzięki wprowadzeniu mechanizmu ograniczającego zmiany w polityce. Może być stosowany do problemów zarówno w środowiskach dyskretnych, jak i ciągłych.

## Inne algorytmy

W powyższych rozdziałach przedstawiono kilka przykładów popularnych algorytmów uczenia ze wzmocnieniem, jednak w omawianej dziedzinie przedstawiono oczywiście bardzo wiele różnych metod, różniących się zastosowaniem i skutecznością.

Do tej pory w pracy skupiono się na metodach, które nie potrzebowały modelu środowiska. Przykładem algorytmu należącego do klasy model-based może być PILCO [27]. Aby modelować zachowanie dynamiczne środowiska wykorzystywane są w nim modele probabilistyczne. Pozwalają one na przewidywanie, jak zmiany w sterowaniu wpłyną na przyszłe stany środowiska. W procesie optymalizacji wykorzystuje metody gradientowe do znalezienia optymalnego sterowania uwzględniając niepewności dynamiczne. Na podstawie kolejnych doświadczeń modele są aktualizowane poprawiając jakość predykcji.

Innym wartym uwagi algorytmem jest MBMF [28] (Model-Based Model-Free), który łączy dwa różne podejścia do uczenia ze wzmocnieniem, czyli metody oparte o model środowiska z tymi, które są go pozbawione. Model w MBMF wykorzystywany jest do oceny kosztu, a więc planowania, natomiast polityka docelowa jest realizowana przy pomocy metod typu model-free.

# Narzędzia implementacji

## Python

## Tensorflow

## Tf-Agents

## Optuna

## TCLab

# Bibliografia

1. Sutton, R. S., & Barto, A. G. (2018). "Reinforcement Learning: An Introduction." <http://incompleteideas.net/book/RLbook2020.pdf>
2. Berger-Tal O, Nathan J, Meron E, Saltz D (2014) The Exploration-Exploitation Dilemma: A Multidisciplinary Framework. PLoS ONE 9(4): e95693. <https://doi.org/10.1371/journal.pone.0095693>
3. Alexandre dos Santos Mignon, Ricardo Luis de Azevedo da Rocha, An Adaptive Implementation of ε-Greedy in Reinforcement Learning, Procedia Computer Science, Volume 109, 2017, Pages 1146-1151, ISSN 1877-0509, <https://doi.org/10.1016/j.procs.2017.05.431>. (<https://www.sciencedirect.com/science/article/pii/S1877050917311134>)
4. Kan WANG, Jinfeng WANG, Jing XIE, Xingxing MA. Improvement of Temperature Control Performance for Electric Heating Water Tank. Thermal Science [Internet]. 2024 Jan;28(1A):25–37.  
   <https://search-1ebscohost-1com-1000003nk00ab.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=asn&AN=176003236&lang=pl&site=eds-live>
5. Khodadadi H, Dehghani A. Fuzzy logic self-tuning PID controller design based on smith predictor for heating system. 2016 16th International Conference on Control, Automation and Systems (ICCAS), Control, Automation and Systems (ICCAS), 2016 16th International Conference on [Internet]. 2016 Oct 1;161–6.   
   <https://search-1ebscohost-1com-1000003nk00ab.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=edseee&AN=edseee.7832314&lang=pl&site=eds-live>
6. Chakrabarti A. , Sadhu P. , Chakraborty A. , & Pal P.. A fuzzy pid controller for induction heating systems with llc voltage source inverter. International Journal of Power Electronics and Drive Systems (IJPEDS) 2017;8(3):1168. <https://doi.org/10.11591/ijpeds.v8.i3.pp1168-1175>
7. Ahmed Sabah Al-Araji. Applying Cognitive Methodology in Designing On-Line Auto-Tuning Robust PID Controller for the Real Heating System. Journal of Engineering [Internet]. 2023 Jul 1;20(09). Available from: <https://search-1ebscohost-1com-1000003nk00ab.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=edsdoj&AN=edsdoj.917559b07303475a820781ef91148d4f&lang=pl&site=eds-live>
8. Chojecki A, Ambroziak A, Borkowski P. Fuzzy Controllers Instead of Classical PIDs in HVAC Equipment: Dusting off a Well-Known Technology and Today’s Implementation for Better Energy Efficiency and User Comfort. Energies (19961073) [Internet]. 2023 Apr;16(7):2967.   
   <https://search-1ebscohost-1com-1000003nk00e2.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=asn&AN=163043447&lang=pl&site=eds-live>
9. Ardabili S. , Beszedes B. , Nádai L. , Széll K. , Mosavi A. , & Felde I.. Comparative analysis of single and hybrid neuro-fuzzy-based models for an industrial heating ventilation and air conditioning control system. 2020 RIVF International Conference on Computing and Communication Technologies (RIVF) 2020. <https://doi.org/10.1109/rivf48685.2020.9140753>
10. Kleiminger W. , Santini S. , & Mattern F.. Smart heating control with occupancy prediction. Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication 2014. <https://doi.org/10.1145/2638728.2641555>
11. Thilker C. , Bergsteinsson H. , Bacher P. , Madsen H. , Calı̀ D. , & Junker R.. Non-linear model predictive control for smart heating of buildings. E3S Web of Conferences 2021;246:09005. <https://doi.org/10.1051/e3sconf/202124609005>
12. Makram Abdellatif, Julien Chamoin, Jean-Marie Nianga, Didier Defer, A thermal control methodology based on a machine learning forecasting model for indoor heating, Energy and Buildings, Volume 255, 2022, 111692, ISSN 0378-7788,  
    <https://doi.org/10.1016/j.enbuild.2021.111692>.
13. Minjae Shin, Sungsoo Kim, Youngjin Kim, Ahhyun Song, Yeeun Kim, Ha Young Kim, Development of an HVAC system control method using weather forecasting data with deep reinforcement learning algorithms, Building and Environment, Volume 248, 2024, 111069, ISSN 0360-1323,  
    <https://doi.org/10.1016/j.buildenv.2023.111069>.
14. Liu B. , Akçakaya M. , & McDermott T.. Automated control of transactive hvacs in energy distribution systems. IEEE Transactions on Smart Grid 2021;12(3):2462-2471. <https://doi.org/10.1109/tsg.2020.3042498>
15. Blad, C., Bøgh, S., & Kallesøe, C. S. (2021). A multi-agent reinforcement learning approach to price and comfort optimization in hvac-systems. Energies, 14(22), 7491. <https://doi.org/10.3390/en14227491>
16. Adjei P, Tasfi N, Gomez-Rosero S, Capretz MAM. Safe Reinforcement Learning for Arm Manipulation with Constrained Markov Decision Process. Robotics [Internet]. 2024 Apr 1;13(4).   
    <https://search-1ebscohost-1com-1000003w3012f.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=edselc&AN=edselc.2-52.0-85191607938&lang=pl&site=eds-live>
17. Bumsoo Park, Alexandra R. Rempel, Alan K.L. Lai, Julianna Chiaramonte, Sandipan Mishra, Reinforcement Learning for Control of Passive Heating and Cooling in Buildings⁎⁎This work was funded by National Science Foundation CBET-1804218., IFAC-PapersOnLine, Volume 54, Issue 20, 2021, Pages 907-912, ISSN2405-8963, <https://doi.org/10.1016/j.ifacol.2021.11.287>.
18. Watkins, C.J.C.H. (1989). Learning from Delayed Rewards (Ph.D. thesis). University of Cambridge. EThOS uk.bl.ethos.330022, <https://www.cs.rhul.ac.uk/~chrisw/new_thesis.pdf>
19. Watkins C. and Dayan P.. Q-learning. Machine Learning 1992;8(3-4):279-292. <https://doi.org/10.1007/bf00992698>
20. Wiering M, van Otterlo M. Reinforcement learning: State-of-the-art. Springer Science & Business Media; 2012.
21. Williams RJ. Simple statistical gradient-following algorithms for connectionist reinforcement learning. Machine Learning [Internet]. 1992 May 1;8(3–4):229–56. <https://search-1ebscohost-1com-1000003bj00bd.wbg2.bg.agh.edu.pl/login.aspx?direct=true&db=edssjs&AN=edssjs.BABE2B18&lang=pl&site=eds-live>
22. X. Wang et al., "Deep Reinforcement Learning: A Survey," in IEEE Transactions on Neural Networks and Learning Systems, vol. 35, no. 4, pp. 5064-5078, April 2024, doi: 10.1109/TNNLS.2022.3207346.
23. Silver, D., Lever, G., Heess, N., Degris, T., Wierstra, D. &amp; Riedmiller, M.. (2014). Deterministic Policy Gradient Algorithms. Proceedings of the 31st International Conference on Machine Learning, in Proceedings of Machine Learning Research 32(1):387-395   
    https://proceedings.mlr.press/v32/silver14.html.
24. Lillicrap, Timothy & Hunt, Jonathan & Pritzel, Alexander & Heess, Nicolas & Erez, Tom & Tassa, Yuval & Silver, David & Wierstra, Daan. (2015). Continuous control with deep reinforcement learning. CoRR. <https://doi.org/10.48550/arXiv.1509.02971>
25. Schulman J, Wolski F, Dhariwal P, Radford A, Klimov O. Proximal Policy Optimization Algorithms. 2017;   
    <https://doi.org/10.48550/arXiv.1707.06347>
26. Scott Fujimoto, Herke van Hoof, and David Meger. (2018). Addressing Function Approximation Error in Actor-Critic Methods.  
    <https://doi.org/10.48550/arXiv.1802.09477>
27. Marc Deisenroth and Carl E Rasmussen. Pilco: A model-based and data-efficient approach to policy search. In Proceedings of the 28th International Conference on machine learning (ICML-11), pages 465–472, 2011
28. Bansal S, Calandra R, Chua K, Levine S, Tomlin C. MBMF: Model-Based Priors for Model-Free Reinforcement Learning. 2017;   
    <https://doi.org/10.48550/arXiv.1709.03153>
29. Weiner A, Geise J. Model-based deep reinforcement learning for accelerated learning from flow simulations. Meccanica: An International Journal of Theoretical and Applied Mechanics AIMETA [Internet]. 2024 May 14;1–18. <https://doi.org/10.48550/arXiv.2402.16543>
30. Graciela Lopez-Muniz, Ana Cerezo-Hernández, Félix del Campo, Tomas Ruiz-Albi, Daniel Álvarez, Tania Alvaro-De Castro, et al. Usefulness of Artificial Neural Networks in the Diagnosis and Treatment of Sleep Apnea-Hypopnea Syndrome. 2017 Apr 5   
    <https://doi.org/10.5772/66570>
31. R. Tadeusiewicz i M. Szaleniec, Leksykon sieci neuronowych. Projekt Nauka.

Fundacja na rzecz promocji nauki polskiej.