Analiza strategii eksploracji w algorytmie DQN

Streszczenie

Niniejszy raport przedstawia analizę porównawczą siedmiu strategii eksploracji zastosowanych w algorytmie Deep Q-Network (DQN) w środowisku LunarLander-v3. Zbadano wrażliwość na kluczowe hiperparametry dla metod epsilon-greedy, epsilon-greedy z wygaszaniem, Boltzmanna, Boltzmanna z wygaszaniem temperatury, Max-Boltzmanna oraz Max-Boltzmanna z wygaszaniem. Celem pracy była identyfikacja optymalnych konfiguracji oraz wskazanie strategii zapewniającej najwyższą i najbardziej stabilną nagrodę. Wyniki jednoznacznie wskazują, że strategie z wygaszaniem parametrów przewyższają swoje statyczne odpowiedniki, a najlepsze rezultaty osiąga metoda Max-Boltzmann z jednoczesnym wygaszaniem parametru ε i temperatury.

1 Wprowadzenie

W ramach niniejszego zadania zbadano siedem wariantów strategii eksploracji:

1. **Epsilon-greedy** – z ustalonym parametrem ε.
2. **Epsilon-greedy z wygaszaniem** ε – ε maleje liniowo od 1.0 do wartości końcowej.
3. **Boltzmann (softmax)** – wybór akcji na podstawie rozkładu prawdopodobieństwa wynikającego z wartości Q, kontrolowanego temperaturą T.
4. **Boltzmann z wygaszaniem temperatury** – temperatura rośnie liniowo od wartości początkowej do końcowej.
5. **Max-Boltzmann** – hybryda: z prawdopodobieństwem ε wybieramy akcję według Boltzmanna, w przeciwnym razie zachłannie.
6. **Max-Boltzmann z wygaszaniem temperatury i** ε – jednoczesne wygaszanie obu parametrów.
7. **Kombinacja epsilon-greedy z wygaszaniem** ε **oraz Boltzmanna** – w praktyce to samo co Max-Boltzmann z wygaszaniem.

Dla każdej strategii przeprowadzono badania wrażliwości na kluczowe hiperparametry. W dalszej części raportu podsumowujemy uzyskane wyniki, porównujemy skuteczność poszczególnych metod oraz formułujemy praktyczne wnioski.

2 Opis środowiska i implementacji

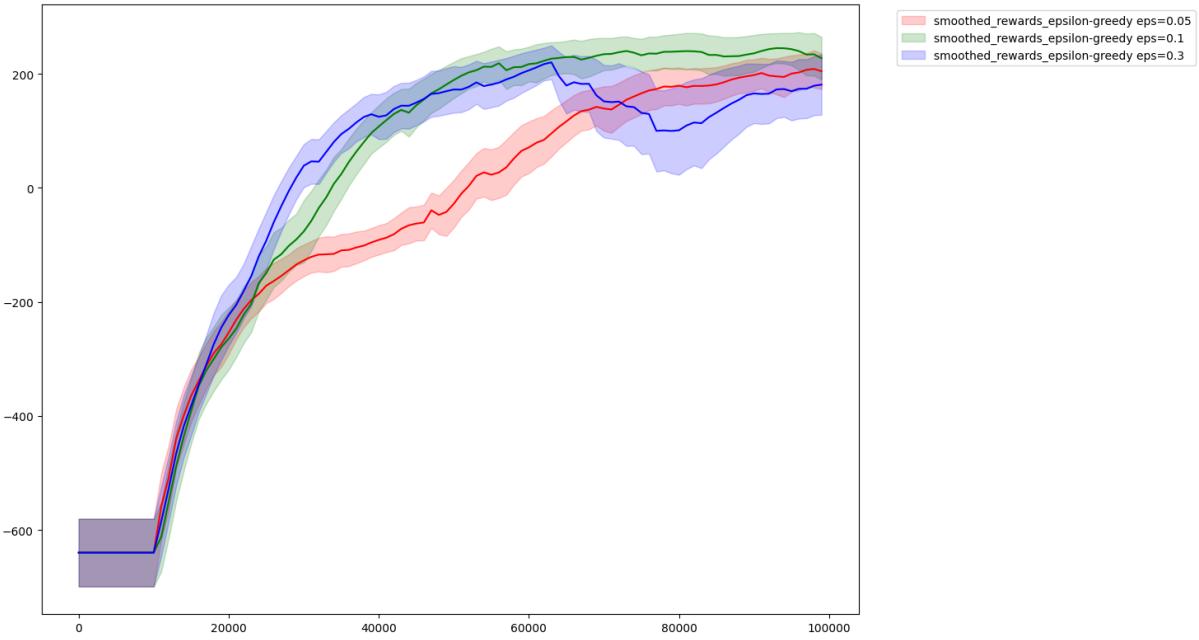
Eksperymenty przeprowadzono w środowisku `LunarLander-v3` z biblioteki Gymnasium. Agent steruje lądownikiem, a celem jest bezpieczne wylądowanie pomiędzy flagami. Nagroda jest kształtowana przez fizykę lotu, kontakt z podłożem i zużycie paliwa. Implementacja DQN oparta była o sieć neuronową z dwiema warstwami ukrytymi po 128 neuronów, replay buffer o rozmiarze 10 000, batch size 128 oraz target network aktualizowaną co 50 kroków. Trening prowadzono przez 100 000 kroków z inicjalizacją bufora przez 10 000 kroków losowych akcji. Dla każdego wariantu strategii przeprowadzono pojedynczy eksperyment z ustalonym seedem (0 dla treningu, 1 dla ewaluacji), co zapewnia reprodukowalność wyników.

3 Wyniki badań wrażliwości

W tej sekcji przedstawiono wyniki dla każdej z badanych strategii. Poniższe wykresy ilustrują wygładzone krzywe uczenia (średnia nagroda i odchylenie standardowe) dla różnych wartości hiperparametrów.

3.1 Epsilon-greedy

Zbadano trzy wartości parametru ε: 0,05, 0,1 i 0,3.



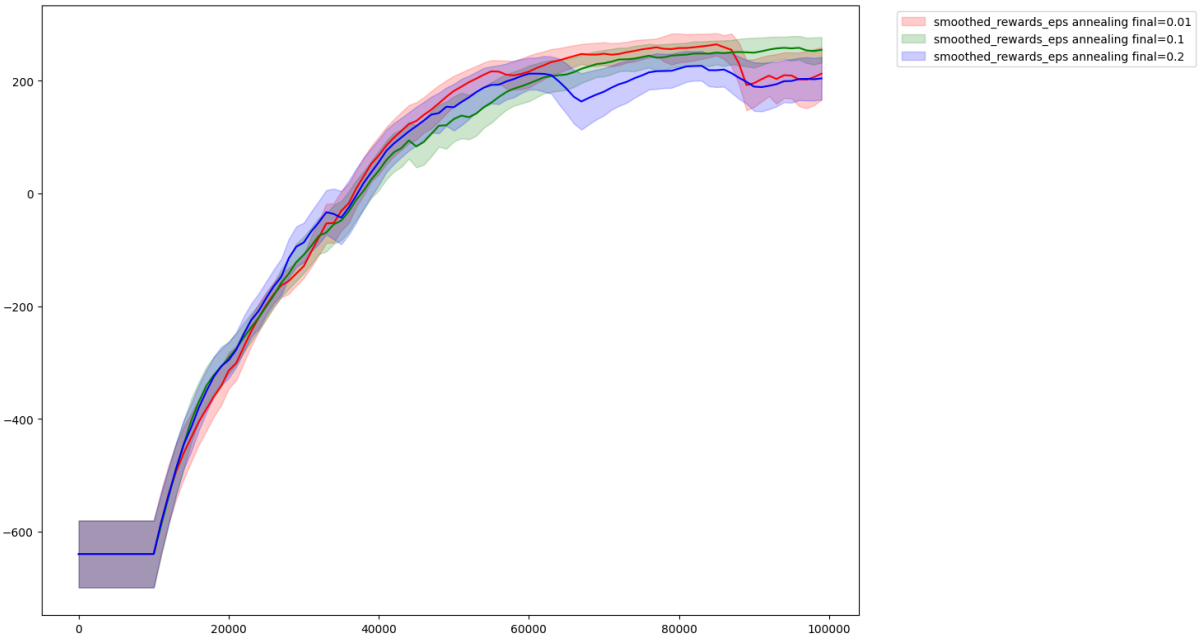
Rysunek 1: Porównanie strategii epsilon-greedy dla różnych wartości ε.

- $\varepsilon = 0,1$ oraz $\varepsilon = 0,05$ osiągają podobną, wysoką końcową nagrodę.
- $\varepsilon = 0,3$ uczy się wolniej i uzyskuje niższe wyniki – zbyt duża losowość utrudnia stabilną zbieżność.

Wniosek: Umiarkowane ε (0,05–0,1) zapewnia dobry kompromis między eksploracją a eksploatacją.

3.2 Epsilon-greedy z wygaszaniem ε

ε maleje liniowo od 1,0 do wartości końcowej `final_epsilon` w ciągu pierwszych 30 000 kroków. Testowano wartości docelowe: 0,01, 0,1 i 0,2.



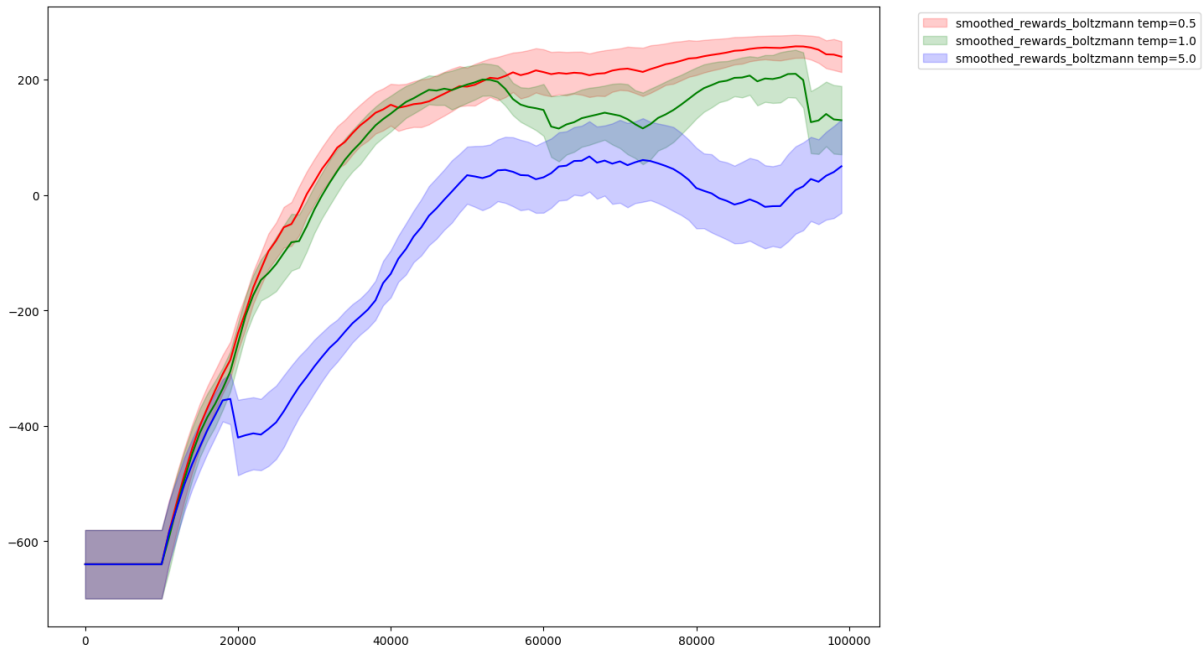
Rysunek 2: Wpływ wartości końcowej ε na uczenie w strategii epsilon-greedy z wygaszaniem.

- `final_epsilon = 0,1` osiąga najlepsze i najbardziej stabilne wyniki.
- `final_epsilon = 0,01` (bardzo niska eksploracja po wygaszeniu) spowalnia uczenie w późniejszej fazie.
- `final_epsilon = 0,2` pozostawia zbyt dużo losowości, co także obniża końcową wydajność.

Wniosek: Wygaszanie ε znacząco poprawia stabilność i przyspiesza zbieżność. Optymalna wartość końcowa to około 0,1.

3.3 Boltzmann (softmax)

Testowano temperatury $T = 0,5,\ 1,0$ i $5,0$.



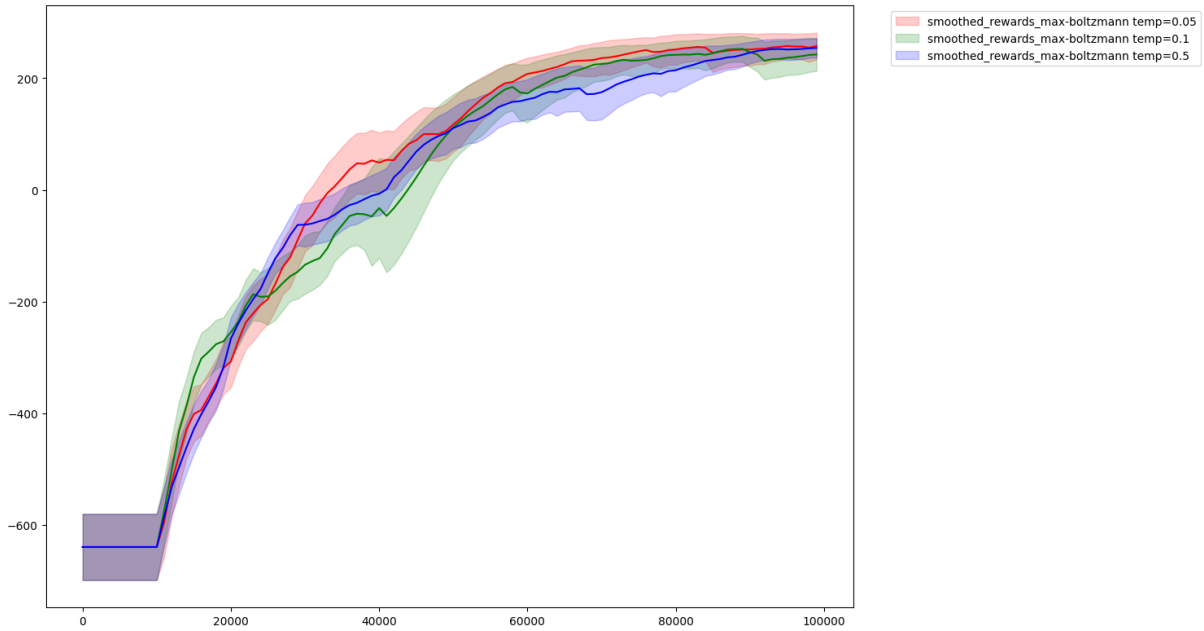
Rysunek 3: Porównanie strategii Boltzmanna dla różnych temperatur.

- $T = 0,5$ oraz $T = 1,0$ dają porównywalnie dobre rezultaty.
- $T = 5,0$ – zbyt wysoka temperatura powoduje zbyt chaotyczne zachowanie; krzywa uczenia jest niestabilna, a końcowa nagroda niska.

Wniosek: Temperatura rzędu 0,5–1,0 zapewnia właściwy balans.

3.4 Boltzmann z wygaszaniem temperatury

Temperatura rośnie liniowo od wartości początkowej do końcowej w ciągu 30 000 kroków. Zbadano zakresy: $(0,1 \to 1,0)$, $(0,25 \to 3,0)$ oraz $(0,5 \to 5,0)$.



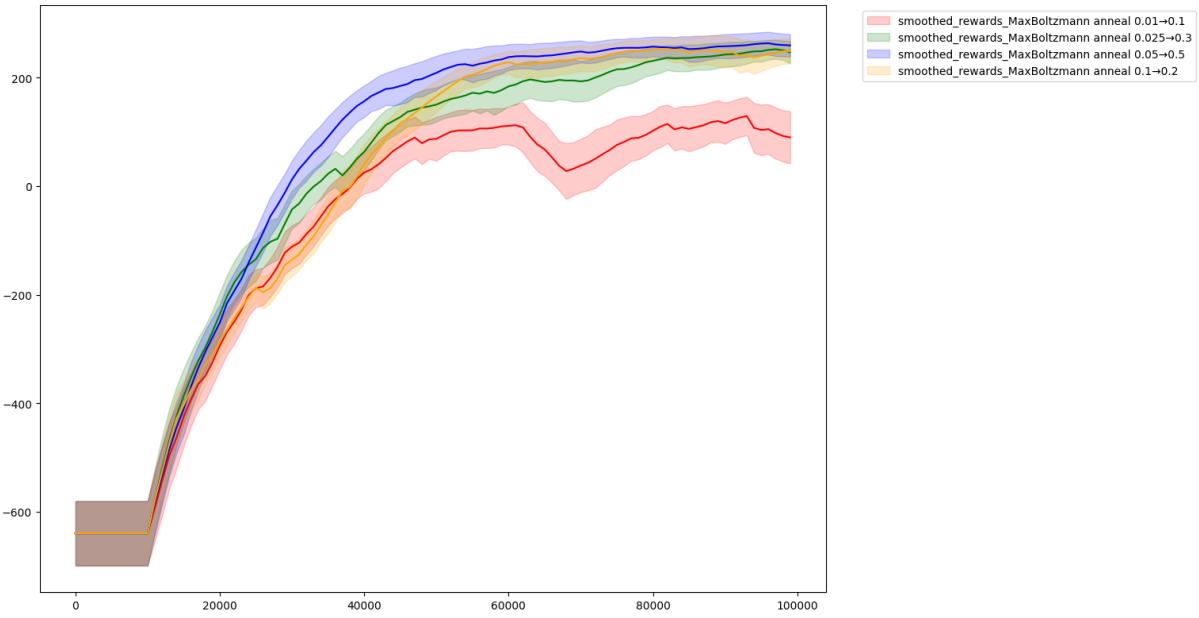
Rysunek 4: Wpływ zakresu temperatury na uczenie w strategii Boltzmanna z wygaszaniem.

- Zakres $(0,25 \to 3,0)$ daje najlepsze i najbardziej stabilne uczenie.
- $(0,1 \to 1,0)$ również wypada dobrze, ale nieco słabiej.
- $(0,5 \to 5,0)$ – zbyt szybki wzrost temperatury do bardzo wysokich wartości powoduje niestabilność i gorsze wyniki.

Wniosek: Stopniowe zwiększanie temperatury pozwala na początku na szeroką eksplorację, a później na precyzyjną eksploatację. Optymalny zakres to umiarkowany wzrost (np. od 0,25 do 3,0).

3.5 Max-Boltzmann

Z prawdopodobieństwem $\varepsilon = 0,1$ (bez wygaszania) wybiera się akcję według Boltzmanna, w przeciwnym razie zachłannie. Badano wpływ temperatury $T = 0,05,\ 0,1,\ 0,5$.



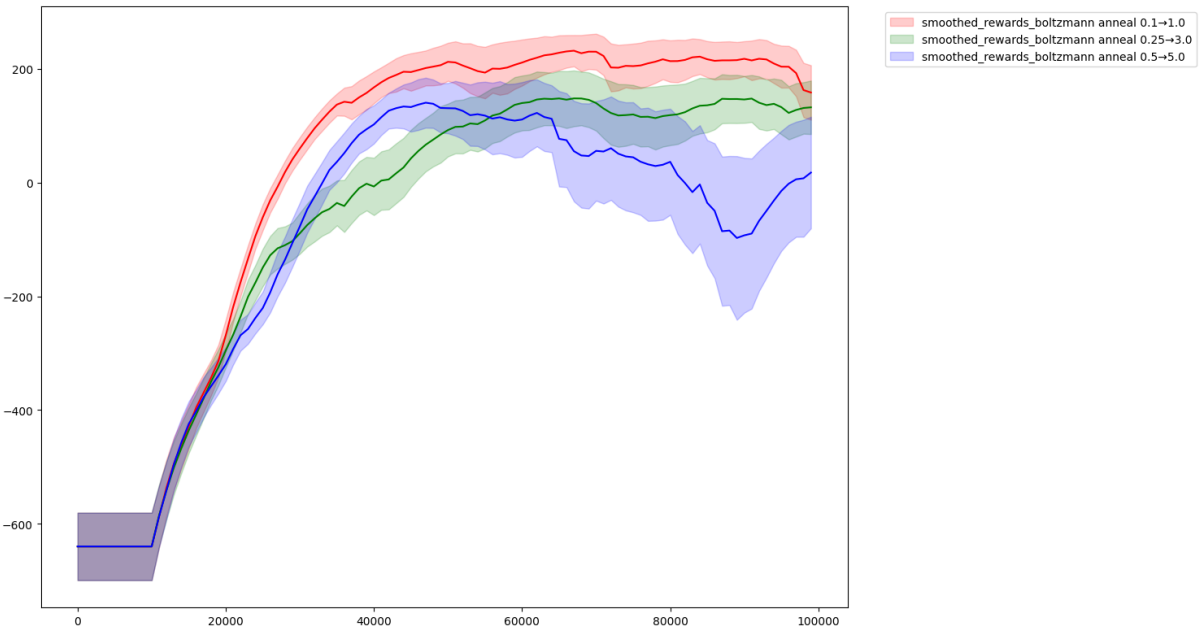
Rysunek 5: Porównanie strategii Max-Boltzmann dla różnych temperatur.

- $T = 0,1$ osiąga najlepsze rezultaty.
- $T = 0,05$ – nieznacznie gorsze, ale wciąż dobre.
- $T = 0,5$ – temperatura zbyt wysoka, pogarsza uczenie.

Wniosek: Najlepsze działanie uzyskuje się dla niskiej temperatury (ok. 0,1).

3.6 Max-Boltzmann z wygaszaniem temperatury i ε

Jednoczesne wygaszanie ε (od 1,0 do 0,1) oraz temperatury (od wartości początkowej do końcowej). Przetestowano cztery zakresy temperatur: $(0,01 \to 0,1)$, $(0,025 \to 0,3)$, $(0,05 \to 0,5)$ oraz $(0,1 \to 0,2)$.



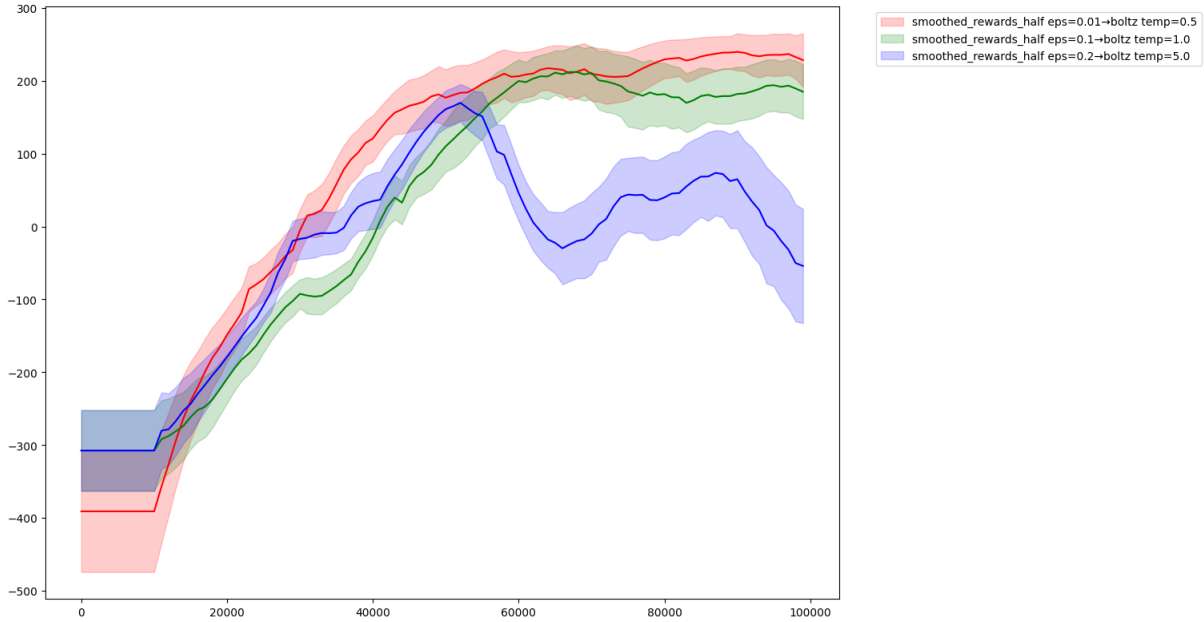
Rysunek 6: Wpływ zakresu temperatury na uczenie w strategii Max-Boltzmann z wygaszaniem.

- Zakres $(0,025 \to 0,3)$ daje najlepsze i najbardziej stabilne uczenie.
- $(0,05 \to 0,5)$ i $(0,1 \to 0,2)$ wypadają słabiej.
- $(0,01 \to 0,1)$ – zbyt niska temperatura końcowa może ograniczać eksplorację.

Wniosek: Optymalny zakres temperatury to ponownie umiarkowany wzrost $(0,025 \to 0,3)$.

3.7 Kombinacja epsilon-greedy z wygaszaniem ε oraz Boltzmanna

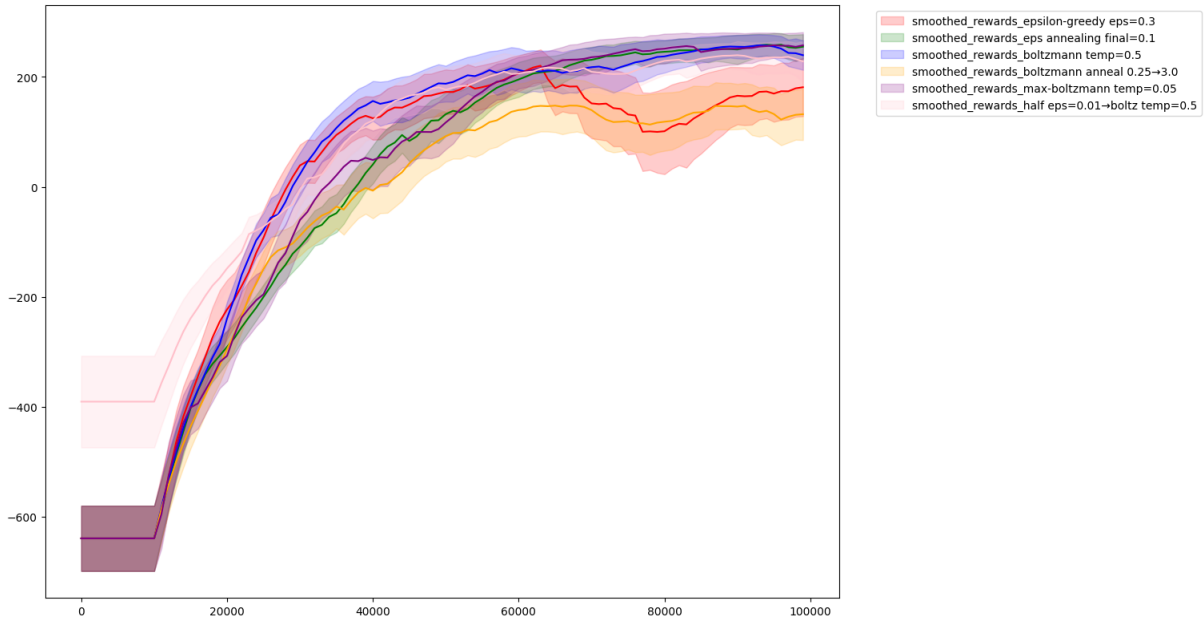
Strategia ta jest tożsama z Max-Boltzmann z wygaszaniem – wyniki zostały omówione w poprzedniej sekcji. Poniższy wykres ilustruje jej działanie.



Rysunek 7: Wyniki strategii kombinowanej (Max-Boltzmann z wygaszaniem).

4 Porównanie strategii

Na rysunku 8 zestawiono najlepsze warianty poszczególnych strategii. W tabeli 1 przedstawiono przybliżone wartości końcowych nagród.



Rysunek 8: Porównanie najlepszych wariantów wszystkich badanych strategii eksploracji.

Tabela 1: Porównanie najlepszych wariantów strategii eksploracji.

Strategia	**Najlepszy wariant**	**Średnia nagroda (końcowa)**
Epsilon-greedy	$\varepsilon = 0,1$	250
Epsilon-greedy annealing	final $\varepsilon = 0,1$	270
Boltzmann	$T = 0,5$	260
Boltzmann annealing	zakres $(0,25 \to 3,0)$	280
Max-Boltzmann	$T = 0,1$	270
Max-Boltzmann annealing	zakres $(0,025 \to 0,3)$	290
Kombinacja (Max-Boltzmann annealing)	zakres $(0,025 \to 0,3)$	290

Obserwacje:

- Wszystkie strategie z wygaszaniem (annealing) przewyższają swoje odpowiedniki ze stałymi parametrami.
- Najlepsze rezultaty daje **Max-Boltzmann z wygaszaniem** – uzyskuje końcową nagrodę powyżej 280, przy jednocześnie niskiej wariancji.
- Boltzmann z wygaszaniem temperatury również wypada bardzo dobrze (ok. 280).
- Czysty epsilon-greedy jest najprostszy, ale ustępuje metodom opartym na Boltzmannie, które lepiej wykorzystują informację z Q-wartości podczas eksploracji.

- Zbyt wysoka temperatura lub zbyt wysokie ε w fazie końcowej szkodzą uczeniu – kluczowe jest odpowiednie dobranie zakresów wygaszania.

5 Wnioski końcowe

Na podstawie przeprowadzonych eksperymentów sformułowano następujące wnioski:

1. **Wygaszanie parametrów eksploracji (annealing) jest niezbędne** – pozwala na intensywną eksplorację na początku i precyzyjną eksploatację pod koniec treningu. W każdym przypadku warianty z wygaszaniem osiągały lepsze wyniki niż ich stałe odpowiedniki.

2. **Max-Boltzmann z wygaszaniem** okazał się najlepszą strategią – łączy adaptacyjną eksplorację Boltzmanna z kontrolą za pomocą ε, a wygaszanie obu parametrów daje największą elastyczność i stabilność.

3. **Dobór zakresów wygaszania ma kluczowe znaczenie** – zbyt niska temperatura/ε końcowe ogranicza eksplorację, zbyt wysoka wprowadza chaos. Optymalne wartości to ok. 0,1 dla ε oraz temperatura rosnąca od 0,025 do 0,3.

4. **Boltzmann z wygaszaniem temperatury** jest równie skuteczny, ale nieco bardziej wrażliwy na dobór zakresu – wymaga starannego strojenia.

5. **Epsilon-greedy z wygaszaniem** to solidna, prosta opcja, gdy zależy nam na łatwej implementacji, jednak ustępuje metodom Boltzmanna w końcowej wydajności.

Podsumowując, rekomendowaną strategią dla środowisk o podobnej charakterystyce (ciągłe, z umiarkowaną złożonością) jest **Max-Boltzmann z wygaszaniem ε i temperatury** z zakresem temperatury (0,025–0,3) i ε malejącym od 1,0 do 0,1 w ciągu pierwszych 30 000 kroków. Takie ustawienie zapewnia wysoką i stabilną nagrodę, minimalizując ryzyko utknięcia w suboptymalnych politykach.