

Rețele neuronale și deep learning

Cursul 10: Reinforcement learning: tranzacționare, acoperirea riscului și dispecerizarea bateriilor



Prof. dr. Daniel Traian PELE

Academia de Studii Economice din București (ASE), Departamentul de Statistică și Econometrie
Academia Română, Institutul de Prognoză Economică (IPE)
IDA Institute Digital Assets ▪ AI4EFin ▪ MSCA Digital Finance

Obiectivele cursului

- Să formulați o decizie economică drept proces decizional Markov (MDP): stare, acțiune, recompensă, tranziție și factor de actualizare.
- Să rezolvați exact un MDP de mici dimensiuni cu ecuația Bellman și programarea dinamică (DP) și să interpretați funcția de valoare.
- Să antrenați agenți bazați pe valoare (Q-learning tabelar, o rețea Q profundă (DQN)) și un agent cu gradient de politică (optimizarea proximală a politicii (PPO)) pe prețuri reale, cu mai multe valori inițiale și un reper corect.
- Să acoperiți riscul unei opțiuni cu o politică neuronală care minimizează pierderea așteptată în coada distribuției (ES) și să o comparați cu delta Black-Scholes.
- Să judecați cînd reinforcement learning (RL) aduce valoare într-o afacere și cînd ajunge o regulă simplă.

Abrevieri

A2C	actor-critic cu avantaj (advantage actor-critic)
AI	inteligență artificială
ASE	Academia de Studii Economice din București
ATM	la bani (prețul de exercitare = prețul curent)
BG	Bulgaria (zonă de ofertare)
CET	ora Europei Centrale
CPU	procesor central
DE-LU	zona de ofertare Germania-Luxemburg
DM	testul Diebold-Mariano
DP	programare dinamică
DQN	rețea Q profundă (deep Q-network)
ENTSO-E	Rețeaua europeană a operatorilor de transport și de sistem pentru energie electrică
ES	pierderea așteptată în coada distribuției (expected shortfall)
EUR	euro
GAE	estimarea generalizată a avantajului
HU	Ungaria (zonă de ofertare)
IPE	Institutul de Prognoză Economică, Academia Română
ISE	Institutul Fraunhofer pentru Sisteme de Energie Solară
KL	divergența Kullback-Leibler
MDP	proces decizional Markov
MLP	perceptron multistrat
MW	megawatt
MWh	megawatt-oră
PPO	optimizarea proximală a politicii (proximal policy optimisation)
ReLU	unitate liniară rectificată
RL	reinforcement learning
RO	România (zonă de ofertare)
SSRN	Social Science Research Network
TD	diferență temporală
TRPO	optimizarea politicii în regiune de încredere
UE	Uniunea Europeană
USD	dolar american
VaR	valoarea la risc (value at risk)
VIX	indicele de volatilitate Cboe (volatilitatea implicită pe 30 de zile a indicelui S&P 500)

Datele reale ale cursului

- Trei probleme, trei seturi de date reale
  - Fiecare cifră de pe slide-uri este calculată de analysis_L10.py și salvată în L10_numbers.json.

Problema	Date și sursă	Împărțire (ordine cronologică)
Dispecerizarea bateriei	Prețuri orare pentru ziua următoare: România (RO), Ungaria (HU), Germania-Luxemburg (DE-LU), Bulgaria (BG); energy-charts.info, al Institutului Fraunhofer pentru Sisteme de Energie Solară (ISE), din datele Rețelei europene a operatorilor de transport și de sistem pentru energie electrică (ENTSO-E)	antrenare 2019–2023, validare 2024, test 1 ianuarie 2025 – 24 octombrie 2026
Tranzacționare	Închiderile zilnice S&P 500 și indicele de volatilitate Cboe (VIX); Yahoo Finance	antrenare 2005–2016, validare 2017–2019, test 2020 – 29 octombrie 2026
Acoperirea opțiunii	Traiectorii S&P 500 de 21 de zile, VIX ca volatilitate implicită	traiectorii începute în 2005–2016 / 2017–2019 / 2020 – 1 octombrie 2026

- Datele se opresc la 31 octombrie 2026, ultima lună completă
  - Valorile inițiale 20261108 ... 20261108+4
  - Antrenare pe procesorul central (CPU) Apple M5 Max, cu rezultatele salvate.

Cuprins

- Decizii secvențiale în afaceri
- Procese decizionale Markov și ecuația Bellman
- Învățarea bazată pe valoare: Q-learning și DQN
- Gradienți de politică și PPO: tranzacționarea indicelui S&P 500
- Acoperirea neuronală: o politică ce minimizează riscul din coadă

Reinforcement learning și învățarea supervizată

- **Pe scurt:** învățarea supervizată prognozează o etichetă; RL alege acțiuni cu efecte întîrziate, care depind de deciziile anterioare.

- **Supervizată**
  - Date $(\mathbf{x}_i, y_i)$, pierderea $\mathcal{L}(\hat{y}_i, y_i)$ cunoscută pentru fiecare exemplu
  - Prognoza nu schimbă exemplul următor
- **Reinforcement learning**
  - Agentul vede o stare s_t, acționează a_t, primește o recompensă r_{t+1}
  - Acțiunea schimbă starea următoare (o baterie plină nu se mai încarcă)
  - Nimeni nu indică acțiunea corectă: doar o recompensă întîrziată și zgomotoasă

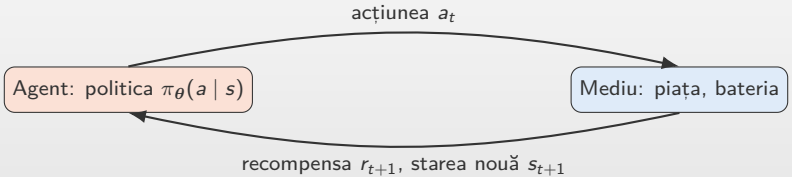
- **Decizii economice de acest tip**
  - Stocare: cînd se încarcă și cînd se descarcă
  - Tranzacționare: cînd se intră și cînd se iese din piață
  - Acoperire: cît din activul suport se deține în fiecare zi
  - Stocuri, prețuri, limite de credit

- **În practică:** folosiți RL doar cînd acțiunea de azi schimbă opțiunile de mîine.

(Sutton & Barto, 2018, ch. 1)

Bucla agent-mediu

- **Pe scurt:** agentul acționează, mediul răspunde cu o recompensă și o stare nouă; a învăța înseamnă a îmbunătăți politica pe baza acestor răspunsuri.



- **Episod**: o zi de livrare, o fereastră de 126 de zile, durata de 21 de zile a unei opțiuni.
- **Fără model**: agentul învață din tranziții eșantionate $(s_t, a_t, r_{t+1}, s_{t+1})$, nu din legea de tranziție.
- **Aici, offline**: toți cei trei agenți învață din date istorice.

- **În practică:** scrieți starea, acțiunea și recompensa înainte de orice antrenare.

(Sutton & Barto, 2018, ch. 3)

Trei aplicații economice ale reinforcement learning

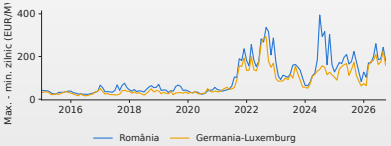
Problema	Starea s_t	Acțiunea a_t	Recompensa r_{t+1}
Baterie	ora, energia stocată, prețurile de ieri	încărcare / inactiv / descărcare	fluxul de numerar al orei (EUR)
Tranzacționare	randamente trecute, volatilitate, VIX, trend, poziția curentă	în afara pieței / long	poziția $\times$ randamentul următor $-$ cost
Acoperire	nivelul indicelui, timpul pînă la scadență, VIX, acoperirea anterioară	raportul de acoperire $\delta_t \in [0, 1]$	rezultatul final, evaluat prin ES

- Aceeași matematică: maximizarea recompensei actualizate așteptate
  - Informație și atitudini față de risc diferite.
- Reperele sînt reguli pe care un manager le folosește deja: planul de ieri, cumpără și păstrează, media mobilă de 200 de zile, delta Black-Scholes.
- **În practică:** un agent RL merită folosit doar dacă bate regula deja folosită, după costuri și pe mai multe valori inițiale.

(calcul propriu)

Venitul bateriilor din diferența zilnică de preț și creșterea acestei diferențe

- **Pe scurt:** o baterie cumpără în orele cele mai ieftine și vinde în cele mai scumpe; venitul ei crește cu diferența zilnică dintre maxim și minim.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: 2 ianuarie 2015 - 31 octombrie 2026, media lunară a diferenței zilnice.

- România: diferența zilnică medie a fost 47,9 EUR/MWh în 2019, 164,3 în 2025 și 182,8 în 2026 pînă acum.
- Luna de vîrf, iulie 2024: 394,4 EUR/MWh
  - Producția solară coboară acum prețurile de la prînz.

- **În practică:** rentabilitatea unei baterii depinde de diferențele viitoare de preț, nu de nivelul prețului.

(calcul propriu)

NN_L10_spread_monthly

Cuprins

- Decizii secvențiale în afaceri
- Procese decizionale Markov și ecuația Bellman
- Învățarea bazată pe valoare: Q-learning și DQN
- Gradienți de politică și PPO: tranzacționarea indicelui S&P 500
- Acoperirea neuronală: o politică ce minimizează riscul din coadă

Cele cinci elemente ale unui proces decizional Markov

- **Pe scurt:** viitorul depinde de trecut doar prin starea și acțiunea curente.

$$\mathcal{M} = (\mathcal{S}, \mathcal{A}, P, R, \gamma)$$

$$P(s' \mid s, a) = \Pr(s_{t+1} = s' \mid s_t = s, a_t = a)$$

$$G_t = \sum_{k=0}^{T-t-1} \gamma^k r_{t+k+1}, \quad \pi^* = \arg\max_{\pi} \mathbb{E}_\pi[G_0]$$

- **unde**
  - $\mathcal{S}, \mathcal{A}$ = stări și acțiuni, P = legea de tranziție, $R(s,a) = \mathbb{E}[r_{t+1} \mid s, a]$
  - $\gamma \in [0,1]$ = factorul de actualizare, G_t = cîștigul cumulat, $\pi(a \mid s)$ = politica.

- **Proprietatea Markov**: starea trebuie să conțină tot ce contează
  - Pentru baterie: ora și energia stocată
  - Pentru tranzacționare: poziția curentă, deoarece ea determină costul tranzacției următoare.
- **Actualizarea**: $\gamma = 1$ pentru episoadele de o zi ale bateriei (24 de pași), $\gamma = 0{,}99$ pentru tranzacționare.
- **Orizont finit**: bateria pornește goală în fiecare zi, iar valoarea ei la miezul nopții este zero.

- **În practică:** o variabilă de stare omisă, de exemplu poziția curentă, face agentul instabil.

(Sutton & Barto, 2018, ch. 3)

Funcțiile de valoare măsoară cît de bună este o stare

- **Pe scurt:** V este recompensa viitoare așteptată dintr-o stare; Q este același lucru după o primă acțiune dată.

$$V^\pi(s) = \mathbb{E}_\pi[G_t \mid s_t = s]$$
$$Q^\pi(s,a) = \mathbb{E}_\pi[G_t \mid s_t = s, a_t = a]$$
$$A^\pi(s,a) = Q^\pi(s,a) - V^\pi(s)$$

- **unde**
  - A^π = avantajul acțiunii a față de acțiunea medie a politicii
  - Politica optimă este lacomă în raport cu Q^*: $\pi^*(s) = \arg\max_a Q^*(s,a)$.

- **Metodele bazate pe valoare** (Q-learning, DQN) estimează Q^* și acționează lacom.
- **Metodele bazate pe politică** (REINFORCE, PPO) parametrizează direct π_θ și folosesc V doar ca reper.
- **Exemplul bateriei**: $V_h(s)$ = numărul de euro care mai pot fi cîștigați azi de la ora h, cu s MWh stocați.

- **În practică:** V evaluează un activ (o baterie încărcată, o poziție deschisă)
  - Q compară decizii.

(Sutton & Barto, 2018, ch. 3–4)

Ecuația Bellman: descompunerea recursivă a funcției de valoare

- **Pe scurt:** valoarea optimă acum = cea mai bună recompensă imediată + valoarea optimă actualizată a stării în care ajungeți.

$$V^*(s) = \max_a \left\{ R(s,a) + \gamma \sum_{s'} P(s' \mid s,a)\, V^*(s') \right\}$$

$$Q^*(s,a) = R(s,a) + \gamma\, \mathbb{E}_{s'} \left[\max_{a'} Q^*(s',a') \right]$$

- **unde**
  - Pentru $\gamma < 1$ membrul drept este o contracție, deci iterarea valorii $V_{k+1} = \mathcal{T} V_k$ converge la punctul fix unic V^*
  - Pe un orizont finit, o singură trecere înapoi dă soluția exactă.

- **Principiul optimalității** (Bellman, 1957): un plan optim rămîne optim din orice stare intermediară.
- **Blestemul dimensionalității**: tabelul are $|\mathcal{S}| \times |\mathcal{A}|$ celule
  - Rețelele neuronale înlocuiesc tabelul cînd starea este mare.
- **Modelul este necesar**: DP folosește P și R
  - Q-learning și DQN înlocuiesc speranța matematică prin eșantioane.

- **În practică:** fiecare algoritm RL din acest curs este o metodă aproximativă de rezolvare a acestei ecuații.

(Bellman, 1957; Sutton & Barto, 2018, ch. 4)

Bateria ca proces decizional Markov

- **Pe scurt:** starea = (ora, energia stocată); acțiunea = încărcare, inactivitate sau descărcare; recompensa = fluxul de numerar al orei.

$$s_{h+1} = s_h - a_h, \quad s_h \in \{0, 1, 2\} \text{ MWh}, \quad a_h \in \{-1, 0, +1\}$$

$$r_h(a_h) = \begin{cases} \eta_d \, p_h & a_h = +1 \ (\textit{vînzare}) \\ -p_h/\eta_c & a_h = -1 \ (\textit{cumpărare}) \\ 0 & a_h = 0 \end{cases}$$

- **unde**
  - p_h = prețul orei h (EUR/MWh), $\eta_c = \eta_d = \sqrt{0{,}85}$ (85 % pe ciclu)
  - 2 MWh, 1 MWh mutat pe oră, goală la 0:00.

- **Restricție**: nu se încarcă atunci cînd este plină și nu se descarcă atunci cînd este goală (mascarea acțiunilor).
- **Date**: 4 292 de zile de livrare din România cu 24 de ore complete, 2 ianuarie 2015 – 31 octombrie 2026
  - Din oct. 2025, ora este media a patru prețuri de 15 minute.
- **Simplificări**: bateria nu influențează prețul
  - Fără cost de degradare, tarife de rețea sau servicii de sistem.

- **În practică:** energia stocată este cea care face decizia secvențială.

(calcul propriu)

Exercițiu practic: notebook §2

Programarea dinamică rezolvă exact ziua bateriei

- **Pe scurt:** se pornește înapoi de la miezul nopții: la fiecare oră se păstrează cea mai bună acțiune pentru fiecare nivel de energie stocată.

$$V_{24}(s) = 0$$

$$V_h(s) = \max_{a:\, 0 \le s-a \le 2} \left\{ r_h(a) + V_{h+1}(s-a) \right\}, \quad h = 23, \ldots, 0$$

- **unde**
  - $V_h(s)$ = fluxul de numerar maxim de la ora h la miezul nopții, cu s MWh stocați
  - Venitul zilei este $V_0(0)$
  - Argmax în fiecare (h, s) dă acțiunea optimă.

- **Cost**: $24 \times 3 \times 3$ evaluări pe zi
  - Toate cele 4 292 de zile în cîteva milisecunde, vectorizat pe zile.
- **Previziune perfectă**: DP folosește cele 24 de prețuri ale zilei înainte ca ele să fie cunoscute, deci venitul său este o **limită superioară** pentru orice strategie reală.
- **Reperul naiv**: se calculează planul optim de ieri și se aplică azi.

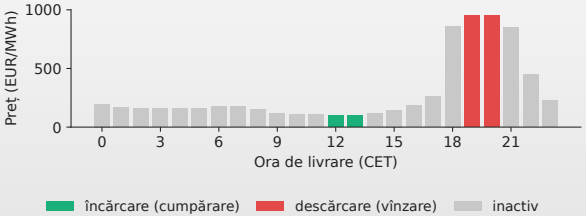
- **În practică:** DP dă plafonul față de care se evaluează orice agent care învață (rata de captare = venit / venitul cu previziune perfectă).

(Bellman, 1957)

Exerciții practic: notebook §2

Programul optim al bateriei pentru o zi

- **Pe scurt:** în ziua cu cea mai mare diferență de preț din 2026, DP încarcă în orele cele mai ieftine și descarcă în cele mai scumpe.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: România, 30 iunie 2026, prețuri orare pentru ziua următoare.

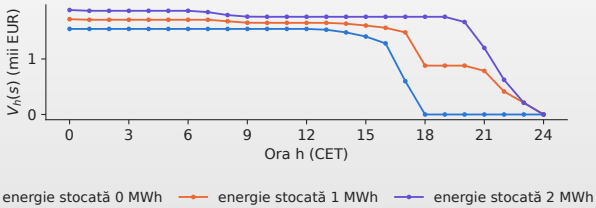
- 30 iunie 2026: prețuri de la 100,65 (ora 13) la 954,59 EUR/MWh (ora 20)
  - 2 ore de încărcare și 2 ore de descărcare.
- Venitul zilei pentru o baterie de 2 MWh: 1 541 EUR.

(calcul propriu)

NN_L10_dp_day

Funcția de valoare evaluează energia stocată oră de oră

- **Pe scurt:** $V_h(2) - V_h(0)$ arată cît valorează doi MWh stocați la ora h: mult înaintea vîrfului de seară, zero după el.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: România, 30 iunie 2026; calcul propriu (inducție înapoi).

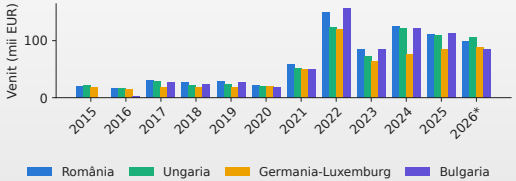
- La ora 12:00, pe 30 iunie 2026: $V_{12}(0) = 1\,541$ EUR, $V_{12}(2) = 1\,760$ EUR, deci un MWh stocat valorează 109,38 EUR, prețul-umbră al energiei din baterie.
- Un Q bine învățat trebuie să reproducă această formă fără să vadă prețurile zilei.

(calcul propriu)

NN_L10_dp_value

Venitul unei baterii de 2 MWh în condiții de previziune perfectă

- **Pe scurt:** venitul DP este o limită superioară; el urmează diferențele de preț, nu nivelul prețului.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: 2015-2026 (*2026: ianuarie-octombrie); zile cu 24 de ore complete.

- România: 28,4 mii EUR în 2019, 150,2 în 2022, 111,2 în 2025, 98,5 în 2026, pe 302 de zile.
- 2025: Ungaria 109,9, Bulgaria 112,9, Germania-Luxemburg 84,3 mii EUR.

- **În practică:** investitorii dimensionează stocarea pe baza acestor cifre
  - Un operator real obține doar o parte, cum se măsoară în secțiunea următoare.

(calcul propriu) NN_L10_dp_revenue_year

Cuprins

- Decizii secvențiale în afaceri
- Procese decizionale Markov și ecuația Bellman
- Învățarea bazată pe valoare: Q-learning și DQN
- Gradienți de politică și PPO: tranzacționarea indicelui S&P 500
- Acoperirea neuronală: o politică ce minimizează riscul din coadă

Q-learning: Q^* învățat din eșantioane, fără model

- **Pe scurt:** $Q(s, a)$ se deplasează spre recompensă plus cea mai bună valoare a stării următoare; diferența este eroarea de diferență temporală (TD).

$$Q(s_t, a_t) \leftarrow Q(s_t, a_t) + \alpha \underbrace{\left[r_{t+1} + \gamma \max_{a'} Q(s_{t+1}, a') - Q(s_t, a_t)\right]}_{\text{TD error } \delta_t}$$

- **unde**
  - α = pasul de învățare (0,05 aici)
  - Ținta folosește acțiunea lacomă din s_{t+1}, indiferent ce face agentul în continuare, deci Q-learning este **off-policy**.
- **Convergența** (Watkins & Dayan, 1992): fiecare pereche (s, a) vizitată de o infinitate de ori, $\sum \alpha_t = \infty$, $\sum \alpha_t^2 < \infty$ (Anexa A).
- **Explorarea**: ε-lacom
  - Acțiune aleatoare cu probabilitatea ε, redusă de la 1 la 0,05.
- **Riscul exploatării**: fără explorare, o estimare proastă de la început nu mai este corectată.

- **În practică:** eroarea TD este un semnal de învățare disponibil după fiecare oră, nu doar la sfîrșitul zilei.

(Watkins & Dayan, 1992)

Exercițiu practic: notebook §3

Definirea stării: informația disponibilă agentului

- **Pe scurt:** operatorul planifică înainte să cunoască prețurile pentru ziua următoare; agentul ține cont de cît de ieftină a fost fiecare oră ieri.

$$s_h = \big(h,\ \text{stored}_h,\ q_h\big), \quad q_h = \left\lfloor \frac{\text{rank}(p_h^{\text{ieri}})}{6} \right\rfloor \in \{0,1,2,3\}$$

- **unde**
  - Rangul orei h între cele 24 de prețuri de ieri, deci $q_h = 0$ = una dintre cele mai ieftine șase ore de ieri
  - Tabelul are $24 \times 4 \times 3$ stări $\times$ 3 acțiuni.

- **Antrenare**: 1 746 de zile din 2019–2023, 40 de treceri
  - Se păstrează epoca cu cel mai mare venit de validare în 2024.
- **Test**: 638 de zile, 1 ianuarie 2025 – 24 octombrie 2026, folosite o singură dată.
- **Scara recompensei**: EUR/100, $\gamma = 1$ în cursul zilei.

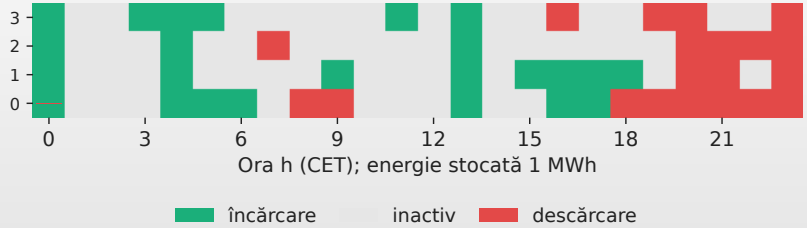
- **În practică:** o stare care include prețurile de azi ar da un test retrospectiv perfect, dar inutil.

(calcul propriu)

Exercițiu practic: notebook §3

Politica învățată prin Q-learning tabelar

- **Pe scurt:** cu 1 MWh stocat, agentul încarcă în orele care au fost ieftine ieri și vinde seara.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: rînduri = quartila prețului orei respective în ziua precedentă; valoarea inițială 20261108.

- O politică ușor de citit: tabelul poate fi tipărit și verificat oră cu oră (288 de stări).
- Ora din zi domină
  - Clasamentul de ieri mută momentul încărcării.

(calcul propriu)

NN_L10_q_policy

Rețeaua Q profundă: aproximarea tabelului Q printr-o rețea neuronală

- **Pe scurt:** tabelul este înlocuit de o rețea $Q_\theta(s,a)$, ajustată prin regresie pe ținta TD.

$$y = r + \gamma\, Q_{\theta^-}\left(s', \arg\max_{a'} Q_\theta(s',a')\right)$$

$$\mathcal{L}(\theta) = \mathbb{E}_{(s,a,r,s')\sim\mathcal{D}}\left[\ell_{\text{Huber}}\left(y - Q_\theta(s,a)\right)\right]$$

- **unde**
  - $\mathcal{D}$ = memoria de reluare (ultimele patru epoci)
  - θ^- = rețeaua-țintă, o copie lentă $\theta^- \leftarrow (1-\tau)\theta^- + \tau\theta$, $\tau = 0{,}005$
  - DQN dublu: rețeaua online alege a', rețeaua-țintă o evaluează.

- **Reluarea** elimină corelația dintre orele consecutive (Mnih et al., 2015).
- **DQN dublu** elimină deplasarea în sus a lui $\max_{a'}$ aplicat unor estimări zgomotoase (van Hasselt et al., 2016).
- **Rețeaua noastră**: 82 de intrări (cele 24 de prețuri standardizate de ieri și din aceeași zi a săptămânii trecute, ziua săptămânii, ora, energia stocată), două straturi de câte 128 de unități liniare rectificate (ReLU), 3 ieșiri.

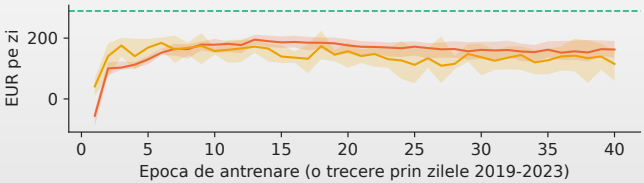
- **În practică:** aproximarea funcției, autoreferința țintei și datele off-policy pot produce divergență
  - Reluarea și rețeaua-țintă mențin stabilitatea.

(Mnih et al., 2015; van Hasselt, Guez & Silver, 2016)

Exercițiu practic: notebook §4

Curbele de învățare: venitul de validare pe epoci

- **Pe scurt:** ambii agenți progresează repede, dar rămîn sub planul de ieri; epoca păstrată este cea mai bună pe 2024, niciodată pe anii de test.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: anul de validare 2024; banda = intervalul celor 5 valori inițiale.

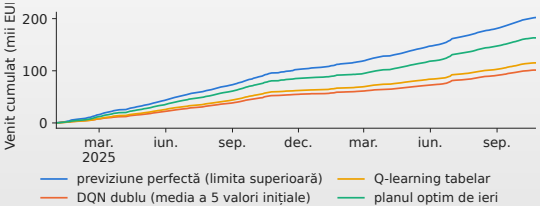
- Cel mai bun venit de validare pe zi, media a 5 valori inițiale: DQN 200 EUR, Q-learning 211 EUR, planul de ieri 289 EUR.

(calcul propriu)

NN_L10_battery_learning

Ponderea venitului cu previziune perfectă obținută în anii de test

- **Pe scurt:** venitul cumulat de la 1 ianuarie 2025 la 24 octombrie 2026; distanța față de linia de sus este valoarea informației perfecte.



Sursa: energy-charts.info (Fraunhofer ISE), din datele burselor și ale ENTSO-E. Date: 1 ianuarie 2025 - 24 octombrie 2026 (test); antrenare 2019-2023, validare 2024.

- Previziune perfectă 202,1
  - DQN 101,3
  - Q-learning 115,2
  - Planul de ieri 163,1 mii EUR.

(calcul propriu)

NN_L10_battery_test_cum

Rezultatele bateriei pe anii de test, cinci valori inițiale

Strategia	Venit (mii EUR)	Captare (%)	Interval (%)
Previziune perfectă (DP, limita superioară)	202,1	100	–
DQN dublu	101,3	50,1	43,9 – 58,5
Q-learning tabelar	115,2	57,0	50,2 – 61,9
Planul optim de ieri (naiv)	163,1	80,7	–

- **DQN față de planul naiv**: statistica Diebold-Mariano (DM) pe diferențele zilnice, variantă Newey-West (7 decalaje): $t = -18{,}99$, $p = 0{,}000$
  - Diferența medie $-96{,}8$ EUR pe zi.
- DQN bate planul naiv pentru 0 din 5 valori inițiale, Q-learning pentru 0 din 5: regula simplă cîștigă, pentru că optimizează pe toate cele 24 de prețuri de ieri.

- **În practică:** raportați rata de captare cu intervalul pe valori inițiale și un test față de planul naiv, niciodată doar cea mai bună rulare.

(Diebold & Mariano, 1995; calcul propriu)

Exercițiu practic: notebook §4

Cuprins

- Decizii secvențiale în afaceri
- Procese decizionale Markov și ecuația Bellman
- Învățarea bazată pe valoare: Q-learning și DQN
- Gradienți de politică și PPO: tranzacționarea indicelui S&P 500
- Acoperirea neuronală: o politică ce minimizează riscul din coadă

Metodele de gradient al politicii

- **Pe scurt:** politica se parametrizează direct și se urcă pe gradientul cîștigului așteptat.

$$\nabla_\theta J(\theta) = \mathbb{E}_{\pi_\theta}\left[\sum_t \nabla_\theta \log \pi_\theta(a_t \mid s_t)\left(G_t - b(s_t)\right)\right]$$

- **unde**
  - $J(\theta) = \mathbb{E}_{\pi_\theta}[G_0]$
  - $\nabla \log \pi =$ funcția scor
  - $b(s_t) =$ reperul, orice funcție de stare, care lasă gradientul nedeplasat și îi reduce varianța
  - $b = V_\phi(s_t)$ dă o metodă actor-critic.

- **REINFORCE** (Williams, 1992): G_t estimat Monte Carlo
  - Nedeplasat, variață mare.
- **Teorema gradientului politicii** (Sutton et al., 2000): gradientul nu cere derivata legii de tranziție, deci piața poate rămîne o cutie neagră.
- **Politicile stocastice** explorează singure și tratează acțiuni continue.

- **În practică:** trucul log-verosimilității este același folosit la antrenarea clasificatorilor, ponderat cu cît de bun a fost rezultatul.

(Williams, 1992; Sutton et al., 2000)

Actor-critic și estimarea avantajului

- **Pe scurt:** un critic V_ϕ prognozează cîștigul, iar actorul învață din surpriza $\hat{A}_t$.

$$\delta_t = r_{t+1} + \gamma V_\phi(s_{t+1}) - V_\phi(s_t), \qquad \hat{A}_t = \sum_{l \geq 0} (\gamma\lambda)^l \, \delta_{t+l}$$

$$\mathcal{L}_V(\phi) = \mathbb{E}\Big[\big(\hat{A}_t + V_\phi^{\text{old}}(s_t) - V_\phi(s_t)\big)^2\Big]$$

- **unde**
  - Estimarea generalizată a avantajului (GAE): $\lambda = 0$ dă avantajul TD de un pas (variantă mică, deplasat), $\lambda = 1$ avantajul Monte Carlo
  - Folosim $\gamma = 0{,}99$, $\lambda = 0{,}95$.

- **În practică:** criticul este o prognoză a recompensei viitoare
  - Un critic slab dă avantaje zgomotoase și o învățare lentă.

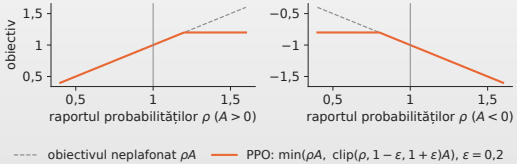
(Schulman et al., 2016)

- **Trunchi comun**: actorul și criticul împart două straturi de cîte 64 de unități tanh
  - Ieșiri: 2 logiți pentru acțiuni și o valoare.
- **Normalizarea avantajului** în fiecare lot păstrează pasul comparabil între regimuri de piață.
- **Bonusul de entropie** 0,01 împiedică politica să devină deterministă prea devreme.

PPO: actualizarea plafonată a politicii

- **Pe scurt:** obiectivul plafonat elimină stimulentul de a muta raportul probabilităților în afara intervalului $[1-\epsilon, 1+\epsilon]$.

$$\mathcal{L}^{\text{CLIP}}(\theta) = \mathbb{E}_t\Big[\min\Big(\rho_t \hat{A}_t,\ \text{clip}(\rho_t, 1-\epsilon, 1+\epsilon)\hat{A}_t\Big)\Big], \quad \rho_t = \frac{\pi_\theta(a_t \mid s_t)}{\pi_{\theta_{\text{old}}}(a_t \mid s_t)}$$



Sursa: Schulman et al. (2017), obiectivul surogat plafonat. Date: funcție analitică, fără date.

- **DATE SIMULATE (doar pentru ilustrare)** Funcție analitică, fără date
  - $\epsilon = 0{,}2$
  - Un înlocuitor ieftin al regiunii de încredere bazate pe divergența Kullback-Leibler (KL) din optimizarea politicii în regiune de încredere (TRPO).

(Schulman et al., 2017)

NN_L10_ppo_clip

Tranzacționarea ca MDP: în afara pieței sau long pe S&P 500

- **Pe scurt:** în fiecare seară agentul alege poziția de mîine; cîștigă randamentul zilei următoare și plătește un cost cînd își schimbă poziția.

$$r_{t+1} = a_t \ln \frac{P_{t+1}}{P_t} - c\,|a_t - a_{t-1}|, \qquad a_t \in \{0, 1\},\ c = 5 \text{ bp}$$

- **unde**
  - $s_t = 7$ variabile standardizate pe 2005–2016 (randamente pe 1, 5, 21, 63 de zile, volatilitatea pe 21 de zile, VIX, distanța față de media de 200 de zile) + a_{t-1}.

- **Antrenare**: 150 de iterații PPO cu cîte 64 de ferestre aleatoare de 126 de zile din 2005–2016, plafon 0,2.
- **Selecție**: iterația cu cel mai bun raport Sharpe în 2017–2019.
- **Test**: 1 716 de zile de la 2 ianuarie 2020, cinci valori inițiale, față de cumpără și păstrează și media mobilă de 200 de zile.

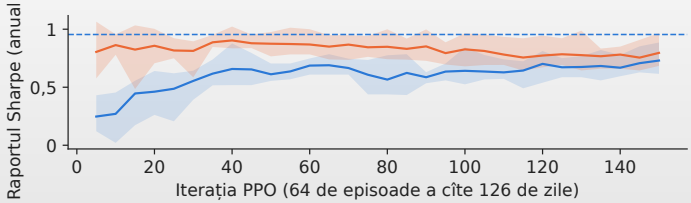
- **În practică:** includeți costul în recompensă, altfel agentul tranzacționează mult prea des.

(Moody & Saffell, 2001; calcul propriu)

Exercițiu practic: notebook §5

Raportul Sharpe la antrenare și validare în timpul antrenării PPO

- **Pe scurt:** raportul Sharpe de antrenare crește; cel de validare este zgomotos, iar cea mai bună iterație este păstrată.



Sursa: Yahoo Finance, cotații zilnice. Date: ferestre de antrenare din 17 octombrie 2005 - 30 decembrie 2016; banda = intervalul celor 5 valori inițiale.

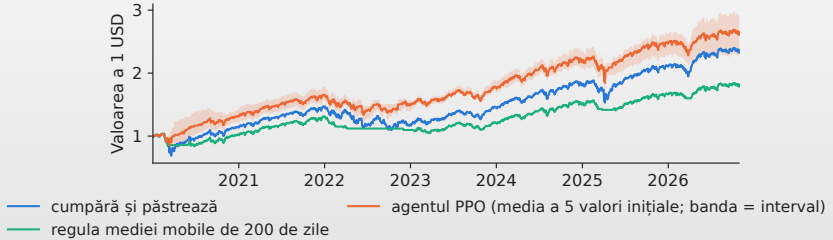
- Raportul Sharpe al strategiei cumpără și păstrează pe validare (2017–2019): 0,95
  - Iterația selectată este în medie 48.

(calcul propriu)

NN_L10_trading_learning

Perioada de test: agentul PPO față de două reguli simple

- **Pe scurt:** valoarea unui USD investit la 2 ianuarie 2020, după un cost de 5 puncte de bază la fiecare schimbare de poziție.



Sursa: Yahoo Finance, cotații zilnice. Date: 2 ianuarie 2020 - 29 octombrie 2026 (test), S&P 500, cost de 5 pb pe tranzacție.

- Raportul Sharpe: cumpără și păstrează 0,61
  - Regula mediei mobile 0,68
  - PPO 0,83 (între 0,72 și 0,95 pe valori inițiale).

(calcul propriu)

NN_L10_trading_equity

Rezultatele tranzacționării pe perioada de test

Strategia	Rand. %/an	Vol. %/an	Sharpe	Scădere max. %	Investit %	Tranz.
Cumpără și păstrează	13,2	20,2	0,61	−33,9	100	1
Media mobilă de 200 de zile	8,9	12,4	0,68	−20,5	79	41
PPO (media a 5 valori inițiale)	15,1	17,0	0,83	−23,2	82	147

- **Incertitudinea**: diferența dintre raportul Sharpe al strategiei PPO mediate pe valori inițiale și cel al strategiei cumpără și păstrează: $+0{,}26$
  - Intervalul de 95 % din bootstrap staționar pe blocuri: $[-0{,}08;\ 0{,}56]$ (2 000 de reeșantionări, bloc mediu de 21 de zile).
- PPO bate strategia cumpără și păstrează la raportul Sharpe pentru 5 din 5 valori inițiale.

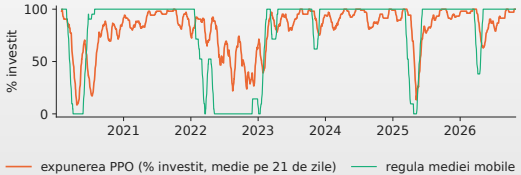
- **În practică:** un interval care conține zero înseamnă că nu există dovezi că agentul aduce valoare, orice ar sugera graficul retrospectiv.

(Politis & Romano, 1994; calcul propriu)

Exercițiu practic: notebook §5

Perioadele în care agentul este în afara pieței

- **Pe scurt:** expunerea agentului PPO (ponderea valorilor inițiale și a zilelor investite) comparată cu regula mediei mobile.



Sursa: Yahoo Finance, cotații zilnice. Date: 2 ianuarie 2020 - 29 octombrie 2026 (test).

- Investit: PPO 82 % din zilele de test
  - 20 feb. – 30 apr. 2020 42 % (media mobilă 14 %)
  - 2022 63 % (media mobilă 15 %).

- **În practică:** citiți politica, nu doar profitul: un agent care este pur și simplu mereu investit este o strategie cumpără și păstrează cu risc de model suplimentar.

(calcul propriu)

NN_L10_trading_exposure

Limitele testelor retrospective pentru agenții de tranzacționare

- **Raport semnal-zgomot redus**: randamentele zilnice ale acțiunilor sînt aproape imprevizibile
  - Agentul învață mai ales zgomotul ferestrelor de antrenare.
- **Nestaționaritatea**: 2005–2016 conține anul 2008
  - 2020–2026 conține prăbușirea din pandemie, șocul inflației și o creștere lungă.
- **Riscul valorii inițiale și al selecției**: cinci valori inițiale dau deja un interval de rapoarte Sharpe
  - Raportarea celei mai bune este supraajustare retrospectivă.
- **Impactul asupra prețului și costurile**: un simulator în care agentul nu influențează prețul ignoră impactul
  - Costurile trebuie incluse în recompensă, nu adăugate ulterior.
- **Unde ajută mai mult RL**: execuția (împărțirea unui ordin mare), crearea de piață, acoperirea riscului, stocarea: probleme cu o stare reală și o recompensă cunoscută.

- **În practică:** cereți distribuția pe valori inițiale, ipoteza de cost și un interval bootstrap înainte de a avea încredere într-un rezultat RL de tranzacționare.

(Hambly, Xu & Yang, 2023)

Cuprins

- Decizii secvențiale în afaceri
- Procese decizionale Markov și ecuația Bellman
- Învățarea bazată pe valoare: Q-learning și DQN
- Gradienți de politică și PPO: tranzacționarea indicelui S&P 500
- Acoperirea neuronală: o politică ce minimizează riscul din coadă

Acoperirea unei opțiuni call vîndute: rezultatul desk-ului

- **Pe scurt:** desk-ul vinde o opțiune call la bani (ATM) pe o lună și tranzacționează zilnic indicele pentru a compensa riscul.

$$\mathrm{PL}(\delta) = C_0 + \sum_{k=0}^{20} \delta_k \, (S_{k+1} - S_k) - (S_{21} - K)^+ - c \sum_{k=0}^{21} S_k \, |\delta_k - \delta_{k-1}|$$

- **unde**
  - S_k = închiderea S&P 500 în ziua k a unei traiectorii reale de 21 de zile, $S_0 = K = 1$
  - C_0 = prima Black-Scholes (VIX din ziua 0, rata 0), δ_k = unitățile deținute pe $(k, k+1]$, $c = 10$ pb.

- **Traiectorii**: una pentru fiecare zi de început
  - Antrenare 3 021 (2005–2016), validare 754, test 1 696 traiectorii din 2020.
- **Test**: prima medie 2,38 %, plata medie 2,53 % din indice (VIX mediu 20,7).
- **Real, nu simulat**: traiectoriile păstrează salturile și grupările de volatilitate din 2008, 2020 și 2022.

- **În practică:** prima abia acoperă plata medie
  - Departamentul de risc gestionează coada.

(Buehler et al., 2019)

Exercițiu practic: notebook §6

Delta Black-Scholes și obiectivul acoperirii neuronale

- **Pe scurt:** reperul acoperă sensibilitatea din model; rețeaua minimizează în schimb pierderea așteptată din coadă a pierderii reale, cu tot cu costuri.

$$\delta_k^{\mathrm{BS}} = \Phi(d_1), \quad d_1 = \frac{\ln(S_k/K) + \frac{1}{2}\sigma_k^2 \tau_k}{\sigma_k \sqrt{\tau_k}}$$

$$\min_{\theta, w} \; w + \frac{1}{1-\alpha}\mathbb{E}\Big[(-\mathrm{PL}(\delta_\theta) - w)^+\Big] = \min_{\theta} \mathrm{ES}_\alpha\Big(-\mathrm{PL}(\delta_\theta)\Big)$$

- **unde**
  - $\sigma_k = \mathrm{VIX}/100$, $\tau_k = (21-k)/252$, $\alpha = 0{,}95$
  - La optim, w este valoarea la risc (VaR)
  - $\delta_\theta(\ln S_k, \tau_k, \sigma_k, \delta_{k-1}) \in [0,1]$: 2 straturi de cîte 32 de unități ReLU, ieșire sigmoidă.

- **Optimizarea politicii cu recompensă cunoscută**: rezultatul este diferențiabil în θ, deci propagăm gradientul înapoi prin toate cele 21 de zile.
- **Acoperirea anterioară ca intrare**: permite rețelei să învețe o bandă fără tranzacții atunci cînd costurile sînt mari.
- **Antrenare**: 600 de pași Adam pe tot setul, cinci valori inițiale.

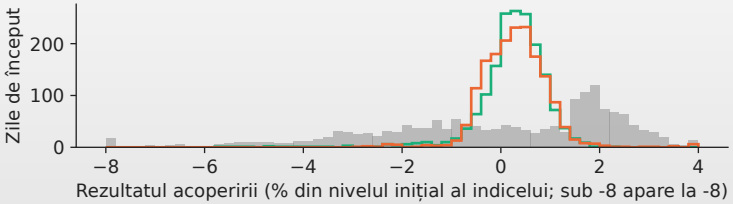
- **În practică:** obiectivul este chiar măsura de risc raportată de desk.

(Buehler et al., 2019; Rockafellar & Uryasev, 2000)

Exercițiu practic: notebook §6

Rezultatul acoperirii pe traiectoriile de test

- **Pe scurt:** acoperirea elimină cea mai mare parte a cozii din stînga; acoperirea neuronală reduce și mai mult coada rămasă, cu un volum tranzacționat mai mic.



Sursa: Yahoo Finance, cotații zilnice. Date: traiectorii de 21 de zile începute 2 ianuarie 2020 - 1 octombrie 2026 (test); S&P 500 și VIX; cost 10 pb.

- Abaterea standard: fără acoperire 2,60, delta 0,74, acoperire neuronală 0,69 (% din indice).

(calcul propriu)

NN_L10_hedge_pnl

Rezultatele acoperirii pe traiectoriile de test, cost de 10 pb

Strategia	PL mediu (%)	Abatere std. (%)	ES_{95} pierdere (%)	Volum tranz.
Fără acoperire	−0,15	2,60	6,82	0
Delta Black-Scholes (VIX)	0,24	0,74	1,94	1,71
Acoperire neuronală (media a 5 valori inițiale)	0,27	0,69	1,15	1,32

- ES_{95} al acoperirii neuronale pe valori inițiale: 1,10 – 1,24
  - Mai mic decît al acoperirii delta pentru 5 din 5 valori inițiale.
- PL = rezultatul (profit sau pierdere) în % din nivelul inițial al indicelui
  - Volumul tranzacționat = media $\sum_k |\delta_k - \delta_{k-1}|$ pe traiectorie.

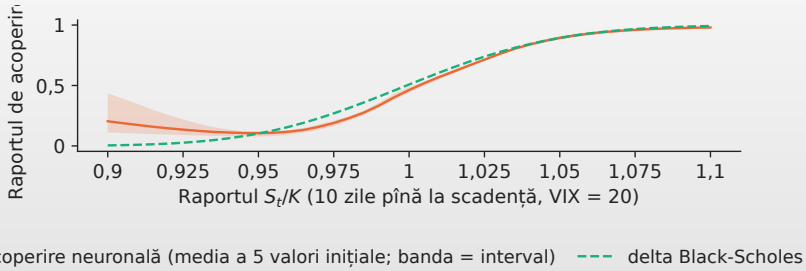
În practică: judecați o acoperire după coada pierderii și după volumul tranzacționat, nu după profitul mediu.

(calcul propriu)

Exercițiu practic: notebook §6

Raportul de acoperire învățat: mai plat și ajustat pentru costuri

- **Pe scurt:** raportul de acoperire în funcție de raportul S_t/K, cu 10 zile înainte de scadență și VIX la 20.



Sursa: Yahoo Finance, cotații zilnice. Date: politică antrenată pe traiectorii de 21 de zile începute în 2005-2016.

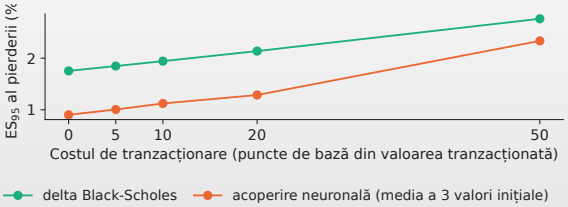
- La bani: acoperirea neuronală 0,46, față de Black-Scholes 0,51
  - Banda arată dispersia pe cinci valori inițiale.

(calcul propriu)

NN_L10_hedge_delta

Costurile de tranzacționare și avantajul acoperirii învățate

- **Pe scurt:** ES_{95} al pierderii din acoperire pentru costuri de la 0 la 50 pb; o rețea antrenată pentru fiecare nivel de cost.



Sursa: Yahoo Finance, cotații zilnice. Date: traiectorii de test începute 2 ianuarie 2020 - 1 octombrie 2026.

- Fără cost: delta 1,75, acoperire neuronală 0,90
  - 50 pb: delta 2,76, acoperire neuronală 2,34 (% din indice).

- **În practică:** valoarea unei acoperiri învățate crește cu fricțiunile: active suport nelichide, diferențe mari între cotații, acoperiri pe energie și mărfuri.

(calcul propriu)

NN_L10_hedge_costs

Reinforcement learning în producție: riscul de model

- **Proiectarea recompensei**: agentul optimizează exact ce scrieți
  - Dacă omiteți costurile, riscul sau restricțiile, va exploata lipsa.
- **Diferența față de simulator**: reluarea istorică ignoră impactul propriu asupra prețului și regimurile viitoare.
- **Evaluarea offline**: o singură perioadă de test, folosită o dată
  - Mai multe valori inițiale
  - Un test față de regula folosită deja.
- **Limite de siguranță**: măștile de acțiune (nivelul de încărcare, limitele de poziție), oprirea de urgență, aprobarea umană.
- **Explicabilitate**: un tabel sau o curbă a raportului de acoperire pot fi auditate
  - Cursul 11 explică politicile neuronale.
- **Monitorizare**: comparați lunar rata de captare și volumul tranzacționat cu testul retrospectiv.

- **În practică:** tratați o politică RL ca pe orice model al băncii: documentată, validată independent, monitorizată și cu o regulă de rezervă.

(Hambly, Xu & Yang, 2023)

Test 1: previziunea perfectă

- Un furnizor arată că agentul său pentru baterie a obținut 95 % din venitul programării dinamice într-un test retrospectiv în care DP a folosit chiar prețurile zilei.
  - Ce întrebați?

Răspuns

- Dacă starea agentului conținea prețurile zilei: dacă da, cei 95 % provin din informație din viitor
  - Un agent corect vede doar ziua de ieri și este comparat cu planul optim de ieri, ca în testul nostru (DQN 50,1 %, planul naiv 80,7 %).

Test 2: plafonarea PPO

- Într-o actualizare PPO, o acțiune are avantajul $\hat{A}_t > 0$, iar noua politică îi dă deja de 1,5 ori probabilitatea veche.
  - Ce gradient transmite obiectivul plafonat pentru acest eșantion, cu $\epsilon = 0{,}2$?

Răspuns

- Zero: $\rho_t = 1{,}5 > 1 + \epsilon$, minimul alege termenul plafonat $1{,}2\,\hat{A}_t$, care nu depinde de $\boldsymbol{\theta}$
  - Politica nu mai este împinsă în acest pas.

Test 3: interpretarea unui rezultat de tranzacționare

- Agentul PPO are, pentru cea mai bună valoare inițială, un raport Sharpe de test peste strategia cumpără și păstrează.
  - Intervalul bootstrap al diferenței de raport Sharpe pentru strategia mediată pe valori inițiale este $[-0{,}08;\ 0{,}56]$.
  - Îl folosiți?

Răspuns

- Nu: cea mai bună valoare inițială este un rezultat selectat, iar un interval care conține zero nu arată o valoare adăugată
  - Păstrați regula simplă.

Recapitulare

- **MDP + Bellman**: scrieți starea, acțiunea, recompensa
  - DP rezolvă exact problemele mici și dă plafonul (bateria: 202,1 mii EUR pe anii de test).
- **Metode bazate pe valoare**: Q-learning și DQN învață din prețurile de ieri
  - Rate de captare 57,0 % și 50,1 %, planul naiv 80,7 %.
- **Gradientul politicii**: PPO pe S&P 500 dă un raport Sharpe de 0,83, față de 0,61 pentru strategia cumpără și păstrează, cu un interval bootstrap care conține zero.
- **Acoperirea neuronală**: ES_{95} al pierderii 1,15 %, față de 1,94 % pentru acoperirea delta, la 10 pb.
- **Verdictul pe date reale**: bateria — regula simplă cîștigă
  - Tranzacționarea — Sharpe mai mare, dar nesemnificativ
  - Acoperirea — politica învățată cîștigă pentru fiecare valoare inițială.
- **Urmează**: cursul 11 deschide cutia neagră: explicarea a ceea ce a învățat o rețea neuronală.

Cercetările grupului nostru: stocare, prețuri și învățare automată în finanțele energiei

- **Arbitrajul cu baterii pe piețele pentru ziua următoare**: prognozele probabilistice comune ale vectorului de prețuri orare alimentează decizia de încărcare și descărcare a unei baterii pe piețele din Europa Centrală și de Est
  - Dispecerizarea are nevoie de distribuția comună, nu doar de prețuri punctuale (Andrei, Cepoi, Dumitrescu, Lessmann, Chiriac (Luchian), Pele & Petukhina, 2026, Social Science Research Network (SSRN) 6872381).
- **Patru generații de modele de preț**: o evaluare comparativă de la modelele statistice la învățarea automată, deep learning și modelele transformer pentru prețurile energiei (Andrei, Velev, Toma, Pele & Lessmann, 2024, arXiv 2411.03372) — prognozele de care are nevoie un agent de dispecerizare.
- **Inteligența artificială (AI) în finanțele energiei**: o analiză sistematică a literaturii despre modul în care învățarea automată și reinforcement learning schimbă deciziile de tranzacționare, de risc și de stocare pe piețele de energie (Anagnoste et al., 2025, Energy Strategy Reviews).
- Rezultatele sînt cele din lucrări; cifrele de pe celelalte slide-uri provin din datele noastre.

(Academia de Studii Economice din București (ASE); Academia Română, Institutul de Prognoză Economică (IPE))

Lecturi și resurse suplimentare

- **Sutton & Barto (2018)**, manualul de referință, gratuit online: incompleteideas.net/book.
- **Spinning Up in Deep RL** (OpenAI), derivări clare pentru gradienții de politică și PPO: spinningup.openai.com.
- **Hambly, Xu & Yang (2023)**: o sinteză a aplicațiilor RL în finanțe.
- **Acest curs**: cod, date, notebook-uri — github.com/danpele/NEURAL_BIZ.
- **Încercați singuri**: notebooks/L10_notebook_ro.ipynb reproduce fiecare metodă din acest curs
  - Seminarul adaugă experimente cu randamentul, dimensiunea bateriei și costurile.
- **Acoperirea neuronală**: Buehler et al. (2019), lucrarea din spatele ultimei secțiuni.

Bibliografie (1/2)

- Andrei, A.-V., Cepoi, C. O., Dumitrescu, B. A., Lessmann, S., Chiriac (Luchian), A.-C., Pele, D. T. & Petukhina, A. (2026). Joint probabilistic forecasting for battery arbitrage in day-ahead power markets. SSRN Working Paper 6872381. doi:10.2139/ssrn.6872381
- Anagnoste, S., Andrei, A.-V., Bolovăneanu, V., Cepoi, C.-O., Clodnitchi, R., Cramer, A.-A., Grecu, R.-A., Lessmann, S., Pele, D. T., Petukhina, A. & Strat, V. A. (2025). The role of AI in (re)shaping energy finance: A systematic literature review. *Energy Strategy Reviews*, 61, 101833. doi:10.1016/j.esr.2025.101833
- Andrei, A.-V., Velev, G., Toma, F.-M., Pele, D. T. & Lessmann, S. (2024). Energy price modelling: A comparative evaluation of four generations of forecasting methods. arXiv:2411.03372
- Bellman, R. (1957). *Dynamic Programming*. Princeton University Press.
- Buehler, H., Gonon, L., Teichmann, J. & Wood, B. (2019). Deep hedging. *Quantitative Finance*, 19(8), 1271–1291. doi:10.1080/14697688.2019.1571683
- Diebold, F. X. & Mariano, R. S. (1995). Comparing predictive accuracy. *Journal of Business & Economic Statistics*, 13(3), 253–263. doi:10.1080/07350015.1995.10524599
- Hambly, B., Xu, R. & Yang, H. (2023). Recent advances in reinforcement learning in finance. *Mathematical Finance*, 33(3), 437–503. doi:10.1111/mafi.12382
- Mnih, V. et al. (2015). Human-level control through deep reinforcement learning. *Nature*, 518, 529–533. doi:10.1038/nature14236
- Moody, J. & Saffell, M. (2001). Learning to trade via direct reinforcement. *IEEE Transactions on Neural Networks*, 12(4), 875–889. doi:10.1109/72.935097

Bibliografie (2/2)

- Politis, D. N. & Romano, J. P. (1994). The stationary bootstrap. *Journal of the American Statistical Association*, 89(428), 1303–1313. doi:10.1080/01621459.1994.10476870
- Rockafellar, R. T. & Uryasev, S. (2000). Optimization of conditional value-at-risk. *Journal of Risk*, 2(3), 21–41. doi:10.21314/JOR.2000.038
- Schulman, J., Moritz, P., Levine, S., Jordan, M. & Abbeel, P. (2016). High-dimensional continuous control using generalized advantage estimation. ICLR. arXiv:1506.02438
- Schulman, J., Wolski, F., Dhariwal, P., Radford, A. & Klimov, O. (2017). Proximal policy optimization algorithms. arXiv:1707.06347
- Sutton, R. S. & Barto, A. G. (2018). *Reinforcement Learning: An Introduction*, 2nd ed. MIT Press. incompleteideas.net/book
- Sutton, R. S., McAllester, D., Singh, S. & Mansour, Y. (2000). Policy gradient methods for reinforcement learning with function approximation. *Advances in Neural Information Processing Systems*, 12, 1057–1063.
- van Hasselt, H., Guez, A. & Silver, D. (2016). Deep reinforcement learning with double Q-learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 30(1). doi:10.1609/aaai.v30i1.10295
- Watkins, C. J. C. H. & Dayan, P. (1992). Q-learning. *Machine Learning*, 8, 279–292. doi:10.1007/BF00992698
- Williams, R. J. (1992). Simple statistical gradient-following algorithms for connectionist reinforcement learning. *Machine Learning*, 8, 229–256. doi:10.1007/BF00992696

Anexa A: convergența Q-learning

- Operatorul de optimalitate Bellman $(\mathcal{T}Q)(s,a) = R(s,a) + \gamma \mathbb{E}_{s'} \max_{a'} Q(s',a')$ este o γ-contracție în norma sup: $\|\mathcal{T}Q_1 - \mathcal{T}Q_2\|_\infty \leq \gamma \|Q_1 - Q_2\|_\infty$.
- Q-learning este o aproximare stocastică a punctului fix: $Q_{t+1} = (1-\alpha_t)Q_t + \alpha_t(\mathcal{T}Q_t + \varepsilon_t)$, cu zgomot ε_t de medie zero.
- Condițiile Robbins-Monro $\sum_t \alpha_t(s,a) = \infty$, $\sum_t \alpha_t^2(s,a) < \infty$ și vizitele infinite dau $Q_t \to Q^*$ aproape sigur (Watkins & Dayan, 1992).
- Cu $\gamma = 1$ pe un orizont finit, argumentul contracției este înlocuit de inducția înapoi pe cei 24 de pași.
- Un $\alpha = 0{,}05$ constant (ca aici) converge doar într-o vecinătate a lui Q^*
  - Setul de validare alege cea mai bună epocă.

(Watkins & Dayan, 1992; Sutton & Barto, 2018, ch. 6)

Anexa B: setările celor trei agenți

Setare	Q-learning / DQN	PPO	Acoperire neuronală
Rețea	tabel; 2×128 ReLU	2×64 tanh, comun	2×32 ReLU, sigmoid
Optimizator	$\alpha = 0{,}05$; Adam $3 \cdot 10^{-4}$	Adam $3 \cdot 10^{-4}$	Adam $3 \cdot 10^{-3}$
Actualizare / orizont	$\gamma = 1$, 24 h	$\gamma = 0{,}99$, $\lambda = 0{,}95$, 126 z	21 de zile
Explorare	ε: $1 \rightarrow 0{,}05$	entropie 0,01	–
Bugetul de antrenare	40 de epoci; 40 de actualizări de 512 pe oră	150 de iterații × 64 de episoade	600 de pași pe tot setul
Selecție	cel mai bun venit în 2024	cel mai bun Sharpe 2017–2019	fără (buget fix)
Timp de rulare (5 valori)	1 838 s	160 s	728 s

(calcul propriu)

Exercițiu practic: notebook §3–§6

Anexa C: estimatorul generalizat al avantajului

- Avantajul pe k pași: $\hat{A}_t^{(k)} = \sum_{l=0}^{k-1} \gamma^l r_{t+l+1} + \gamma^k V(s_{t+k}) - V(s_t) = \sum_{l=0}^{k-1} \gamma^l \delta_{t+l}$.
- GAE este media ponderată exponențial $(1-\lambda)\sum_{k\geq 1} \lambda^{k-1}\hat{A}_t^{(k)} = \sum_{l\geq 0}(\gamma\lambda)^l \delta_{t+l}$.
- Se calculează înapoi într-o singură trecere: $\hat{A}_t = \delta_t + \gamma\lambda\hat{A}_{t+1}$, $\hat{A}_T = 0$ (bucla din analysis_L10.py).
- Compromisul deplasare-varianță: un λ mic are încredere în critic (deplasat dacă V este greșit), un λ mare în recompensele eșantionate (zgomotos).

(Schulman et al., 2016)