Planning and Learning using Adaptive Entropy Tree Search

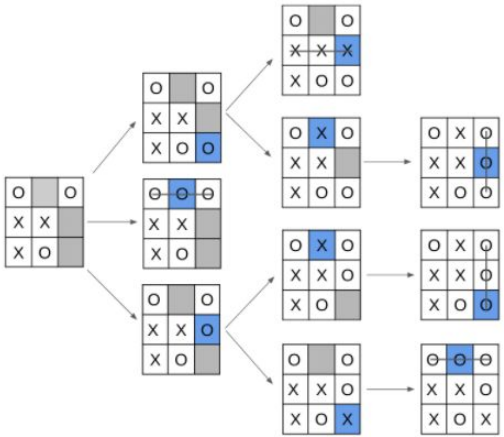
Piotr Kozakowski, Mikołaj Pacek, Piotr Miłoś

Contributions

- We present Adaptive Entropy Tree Search (ANTS) that outperforms MENTS, TENTS and PUCT (AlphaZero) in planning-learning loop setup
- Evaluate planning algorithms in isolation from learning. Our algorithm exhibits superior robustness

Model-Based Reinforcement Learning

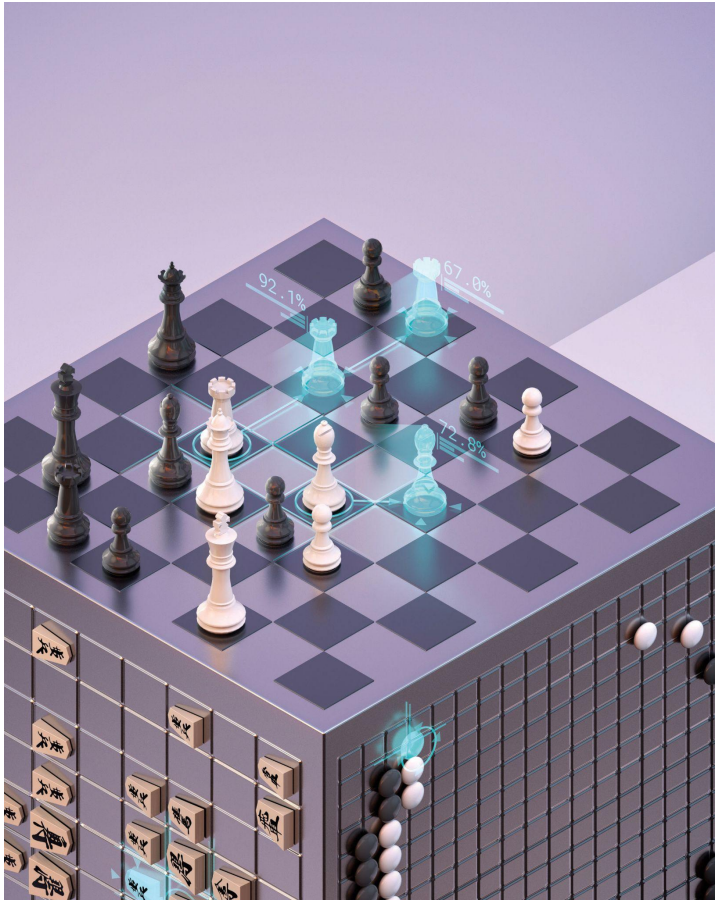
- Planning
- Perfect model of the environment



AlphaZero

Beating world-champions

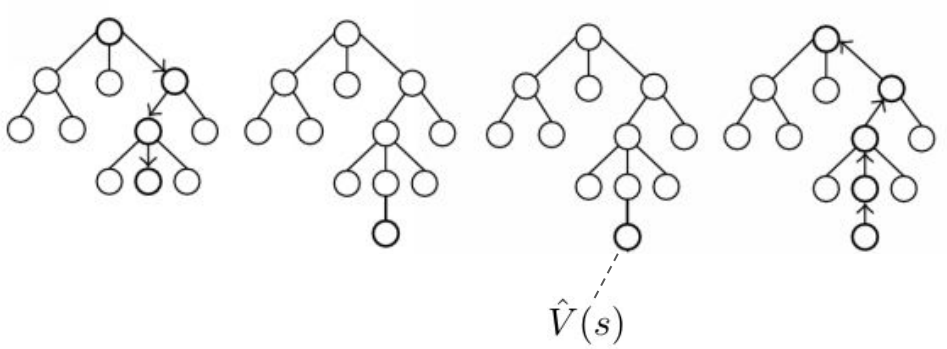
- Chess
- Shogi
- Go



Monte Carlo Tree Search (MCTS)

Classical planning algorithm.

Four phases repeated in a loop:



selection expansion evaluation backup

Monte Carlo Tree Search (MCTS)

Selection using Upper Confidence bounds for Trees (PUCT):

$$\mathrm{PUCT}(s,a) = Q(s,a) + c\hat{\pi}(a|s)\frac{\sqrt{N(s)}}{N(s,a)+1}$$

Backpropagation:

$$V(s) \leftarrow (1-\alpha)\left[\frac{1}{|A|}\sum_{a\in A} Q(s,a)\right] + \alpha\hat{V}(s),\ \alpha = \frac{1}{N(s)}$$

$$Q(s,a) \leftarrow r_{s,a} + \gamma V(s')$$

Schrittwieser et al, 2019 - Mastering Atari, Go, Chess and Shogi by Planning with a Learned Model

Maximum Entropy RL

Exploration by rewarding high-entropy policies

$$J(\pi) = \mathbb{E}_{(s_t, a_t)_{t=0}^{\infty}} \left[\sum_{t=0}^{\infty} \gamma^t \left[\mathcal{R}(s_t, a_t) + \tau \mathcal{H}(\pi(\cdot|s_t)) \right] \right]$$

Ziebart et al, 2018 - Modeling Purposeful Adaptive Behavior with the Principle of Maximum Causal Entropy
Haarnoja et al, 2018 - Soft Actor-Critic: Off-Policy Maximum Entropy Deep Reinforcement Learning with a Stochastic Actor
Eysenbach and Levine et al, 2019 If MaxEnt RL is the Answer, What is the Question?

Maximum Entropy for Tree Search (MENTS)

Selection using the Extended Empirical Exponential Weight (E3W) sampling strategy:

$$\pi_\tau^{\mathrm{e3w}}(a|s) = (1-\lambda_s)\ \pi_\tau(a|s) + \lambda_s \frac{1}{|\mathcal{A}|}, \qquad \lambda_s = \frac{\epsilon|\mathcal{A}|}{\log(N(s)+1)}$$

$$\pi_\tau(a|s) \propto \exp\left(\frac{1}{\tau}Q(s,a)\right),$$

Backpropagation using the optimal entropy-regularized value equation:

$$Q(s,a) \leftarrow \mathcal{R}(s,a) + \gamma \mathbb{E}_{s' \sim \mathcal{P}(\cdot|s,a)} V(s')$$

$$V(s) \leftarrow \tau \log \sum_{a \in \mathcal{A}} \exp\left(\frac{1}{\tau}Q(s,a)\right).$$

Xiao et al, 2019 - Maximum Entropy Monte-Carlo Planning

Dam et al, 2021 - Convex Regularization in Monte-Carlo Tree Search

Adaptive Entropy Tree Search (ANTS)

Temperature adaptation

$$\mathcal{H}^+(\tau) = \left[\frac{1}{|\mathcal{S}_{\text{tree}}|}\sum_{s\in\mathcal{S}_{\text{tree}}}\mathcal{H}(\pi_\tau(\cdot|s))\right] - \mathcal{H}_{\text{avg}},$$
$$\text{where } \pi_\tau(a|s) \propto \exp\left(\frac{1}{\tau}Q(s,a)\right).$$

Pseudoreward shaping

$$Q(s,a;\tau) \leftarrow \mathcal{R}(s,a) + \gamma\mathbb{E}_{s'\sim\mathcal{P}(\cdot|s,a)}V(s';\tau)$$
$$V(s;\tau) \leftarrow \tau\log\sum_{a\in\mathcal{A}}\exp\left(\frac{1}{\tau}Q(s,a;\tau)\right) - \tau\mathcal{H}_{\max},$$

Add temperature to Q-networks' input

Entropies

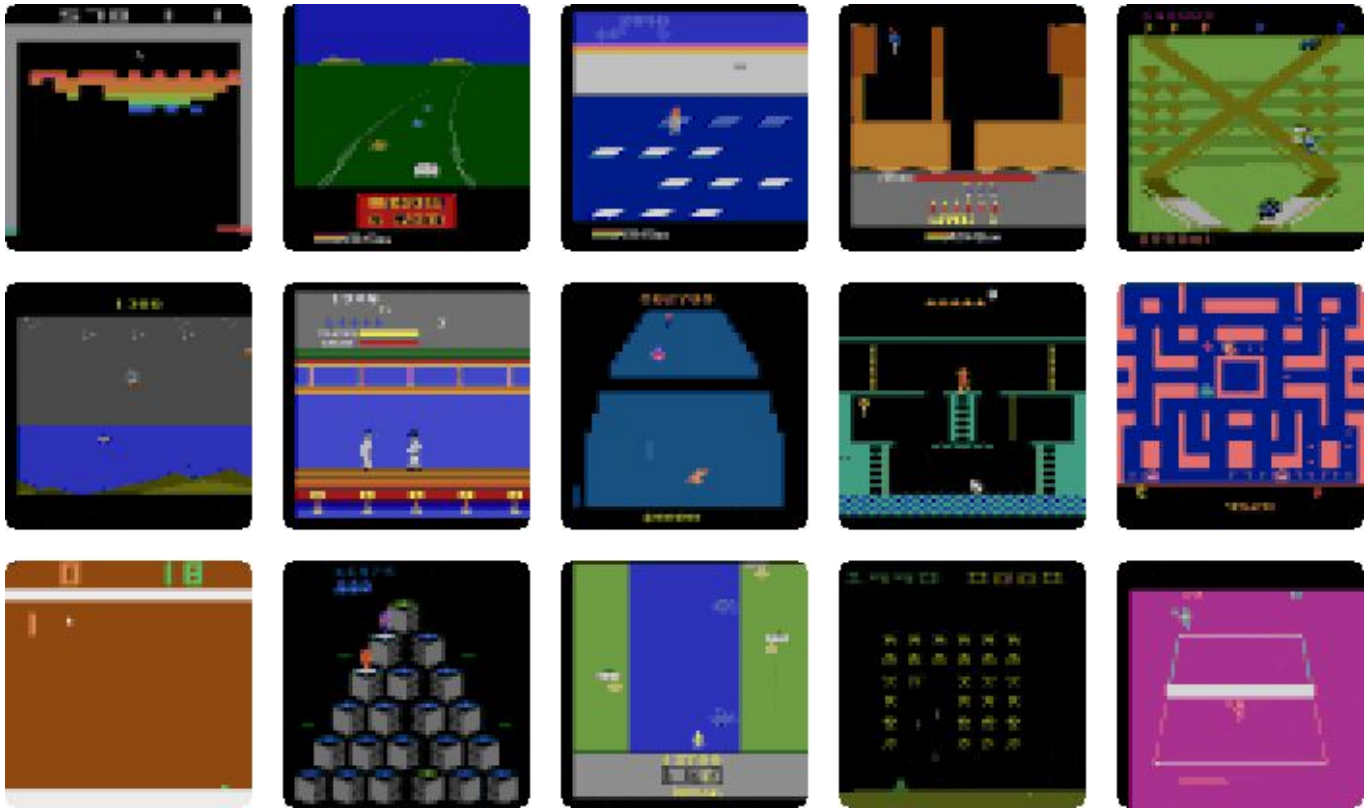
Shannon

$$-\sum_{x\in\mathcal{X}} p(x)\log p(x) = \mathbb{E}[-\log p(X)]$$

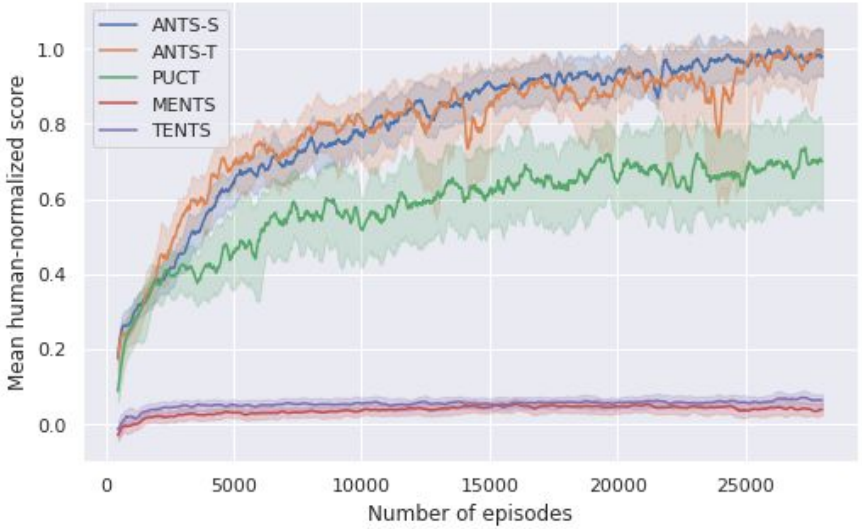
Tsallis

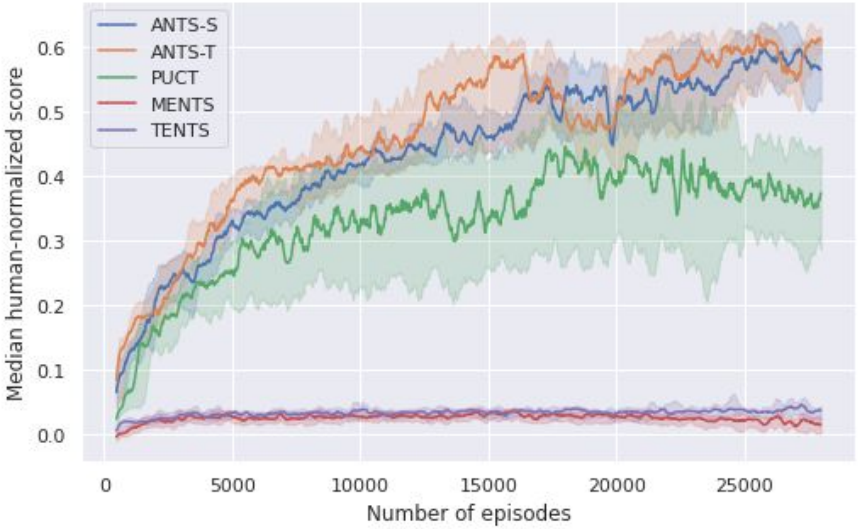
$$\mathbb{E}_{\pi}\left[\frac{1}{2}(1-\pi(a|s))\right]$$

Lee et al, 2017 Sparse Markov Decision Processes with Causal Sparse Tsallis Entropy Regularization for Reinforcement Learning



Planning and learning



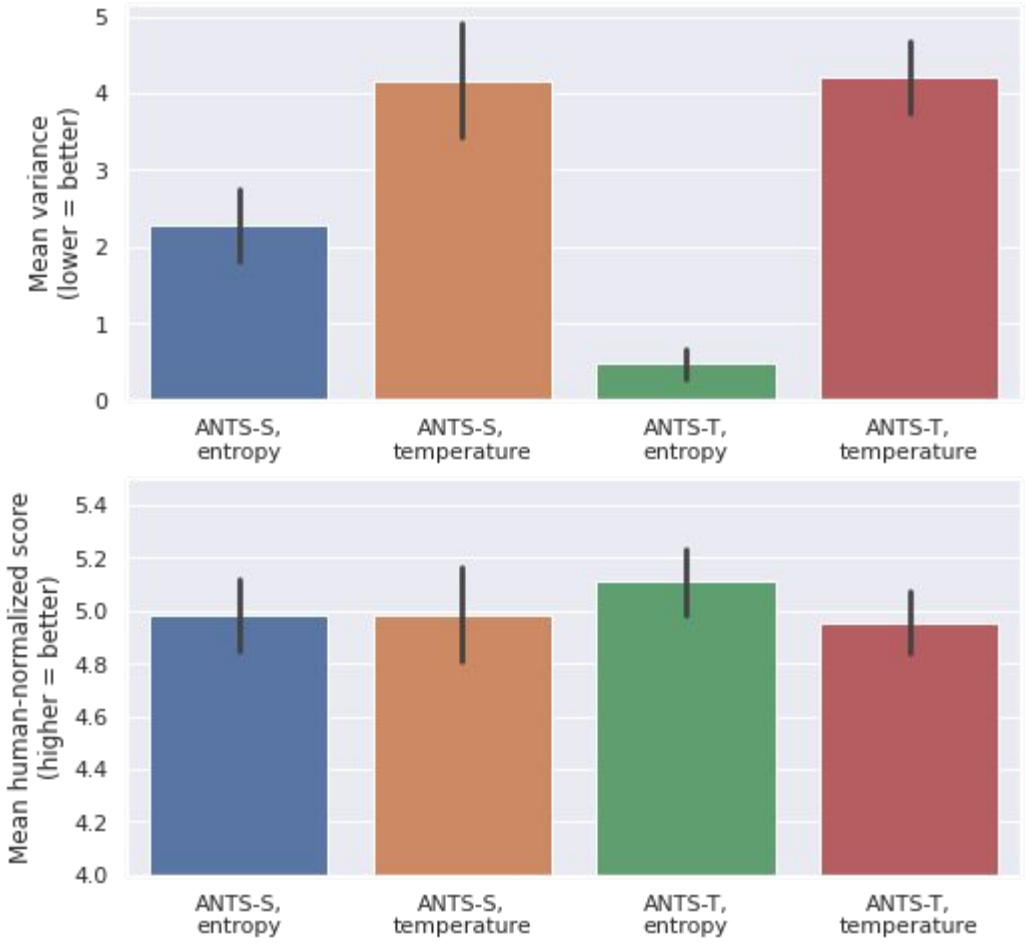


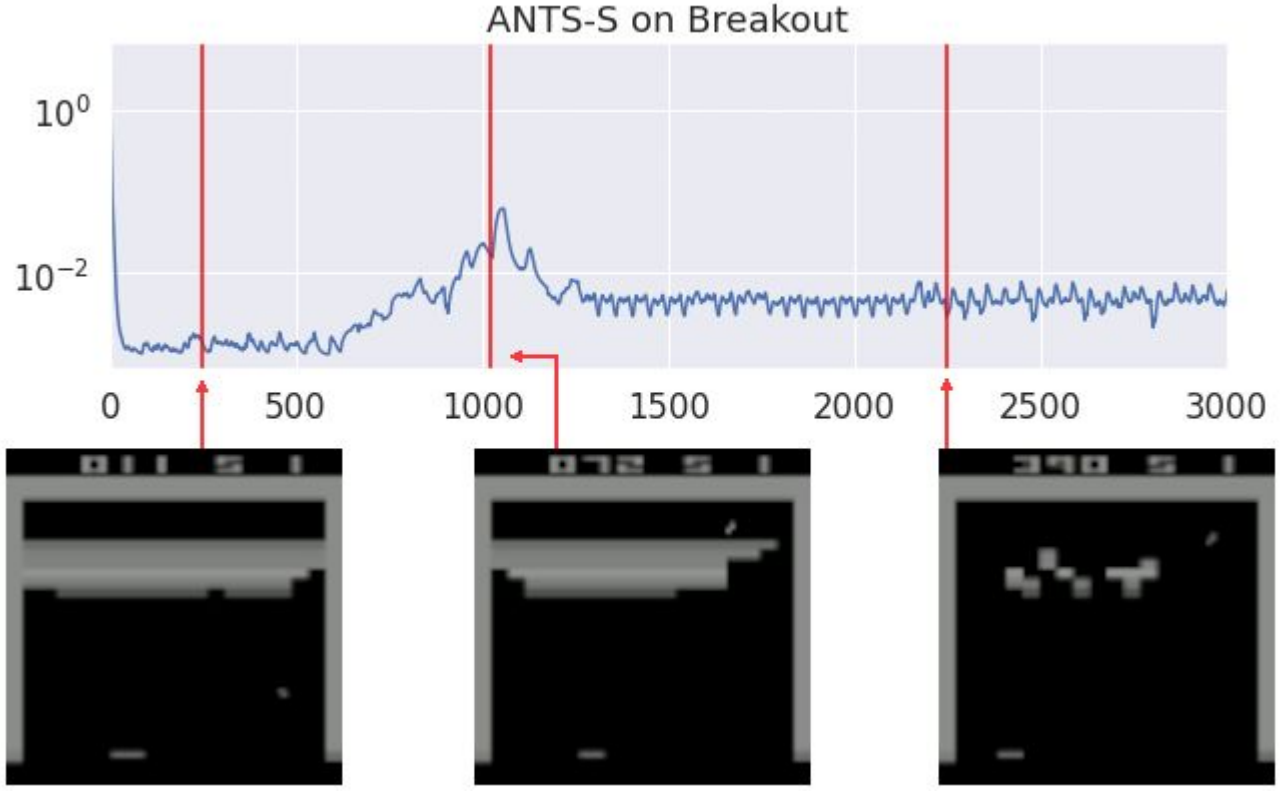
Pretrained Q-networks

	PUCT	MENTS	TENTS	ANTS-S	ANTS-T
Alien	**2853**	**2590**	**2606**	**2783**	2554
Amidar	137	**251**	**236**	148	184
Asterix	**48506**	8305	10406	**51741**	**49758**
Asteroids	**3657**	**3786**	3541	**4059**	**3554**
Atlantis	**278156**	275681	267429	**277870**	**280492**
BankHeist	680	543	472	**707**	**715**
BeamRider	**21745**	9424	7534	**21268**	18536
Breakout	**389**	306	243	351	**377**
Centipede	18662	59167	99631	51387	**133842**
DemonAttack	**55947**	49952	45960	**53998**	**53462**
Enduro	**794**	**800**	**800**	**794**	**800**
Frostbite	**204**	197	197	179	184
Gopher	9606	9330	8150	**10770**	**10766**
Hero	**20715**	19940	19845	20618	20443
MsPacman	3880	3873	3358	**4923**	**4846**
Phoenix	8875	6788	6857	**9187**	**9557**
Qbert	15120	13877	13815	**15472**	15223
Robotank	**55**	27	26	49	51
Seaquest	**3270**	1764	1533	2557	2656
SpaceInvaders	**4630**	2043	2024	**4714**	**4453**
WizardOfWor	**11834**	7063	8030	**12937**	**13386**
#best scores	**14/21**	4/21	3/21	**14/21**	13/21

Robustness

$$\rho(A, H) = \mathbb{E}_{g \sim G} \mathrm{Var}_{h \sim H} \ \texttt{score}(A, h, g)$$




ANTS-S on Breakout
10^0
10^{-2}
0
500
1000
1500
2000
2500
3000

Thanks for the attention