

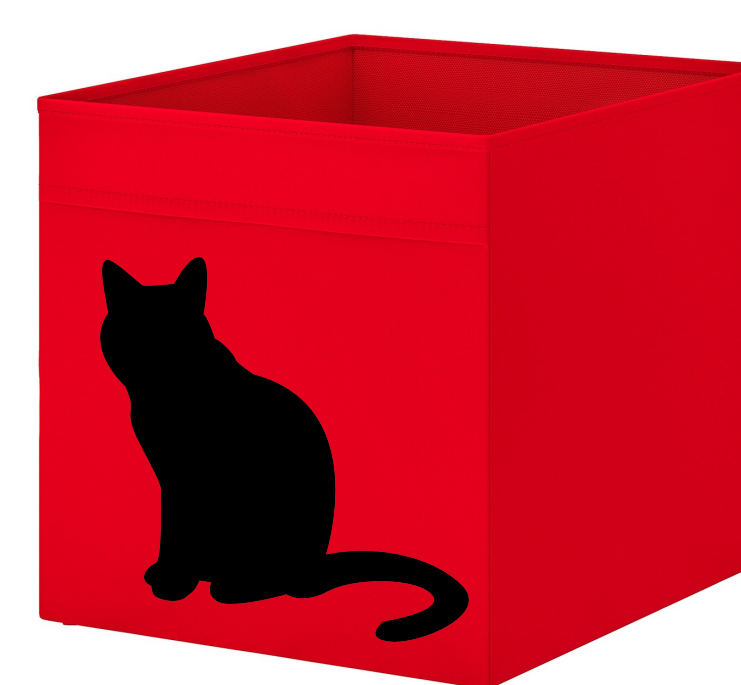
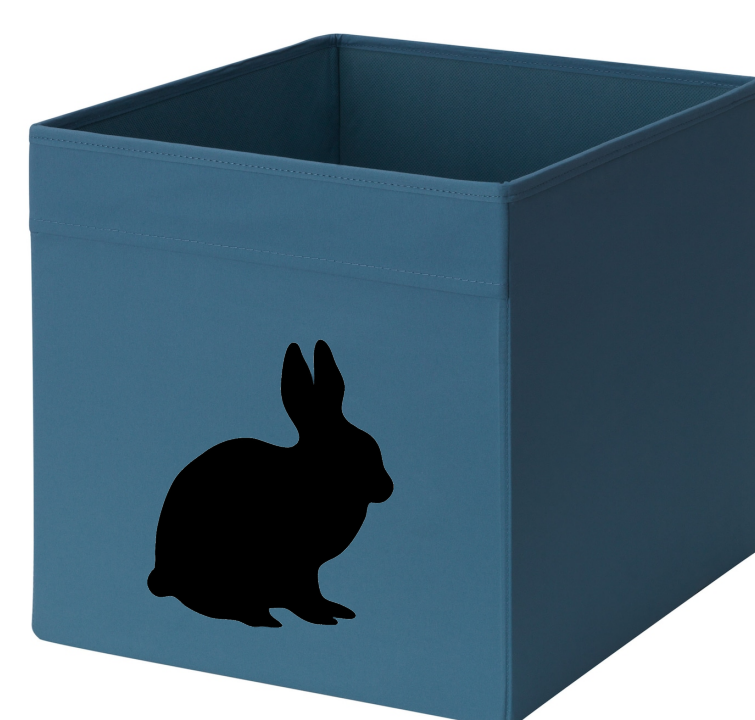
Summary

- We show that the entropy-regularized Bellman operator is **unbiased** when applied to Gaussian Q estimator with inverse-temperature $\beta = \frac{2|\mu_A|}{\sigma_A^2}$
- This β schedule yields a **provably unbiased TD learning** algorithm
- Empirically, such bias reduction **improves convergence**



Why unbiased?

- Bias is like unknown boxes
- Which content is heavier?
- Better scales can't help



Crucial for downstream decisions:

- Exploration
- Further TD updates

Unbiased softmax

Proposition. Let $a \in \{0, 1\}$, and $Q(a)$ a value estimator such that $Q(0) - Q(1) \sim \mathcal{N}(\mu_A, \sigma_A^2)$. For $\beta = \frac{2|\mu_A|}{\sigma_A^2}$, the softmax expectation of Q is an unbiased estimator of the optimal value $\max_a \mathbb{E}[Q(a)]$.

Entropy-regularized Bellman operator

Regularized Bellman operator:

$$\mathcal{B}^{\mathcal{F}}[Q](s, a) = \mathbb{E}[r|s, a] + \gamma \mathbb{E}_{s'|s, a \sim p}[\mathbb{E}_{a'|s' \sim \pi_Q^{\mathcal{F}}}[Q(s', a')]]$$

for

$$\pi_Q^{\mathcal{F}}(\cdot|s) = \operatorname{argmax}_{\pi(\cdot|s)} \mathbb{E}_{a|s \sim \pi}[\mathcal{F}_{Q, \pi}(s, a)]$$

Entropy-regularized objective:

$$\mathcal{F}_{Q, \pi}(s, a) = Q(s, a) - \beta^{-1} \log \pi(a|s)$$

Entropy-regularized Q-learning

Entropy-regularized Q-learning (EQL)

Initialize $Q_1, \dots, Q_k \sim \mathcal{N}(Q^*, \sigma^2)$

for each (s, a, r, s') do

$$(\mu_A, \sigma_A^2) \leftarrow (\text{mean, var})\{Q_i(s', 0) - Q_i(s', 1)\}_{i=1}^k$$

$$\beta \leftarrow \frac{2|\mu_A|}{\sigma_A^2}$$

$$Q_i(s, a) \leftarrow (1 - \alpha)Q_i(s, a) + \alpha(r + \operatorname{softmax}_{a'}(Q_i(s', a'); \beta)) \quad \forall i = 1, \dots, k$$

Experiments

