On the Sample Complexity of (ε, δ)-PAC Learning with the Entropic Risk Measure

Amer Essakine under the supervision of Dr. Claire Vernade

ENS Paris Saclay, MVA master

October 22, 2025

Introduction

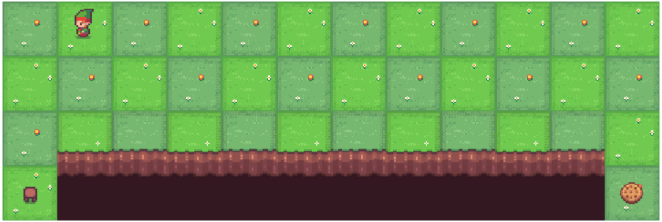


Figure: The cliff environmennt [6]

Markov decision process

- Consider an finite episodic MDP $\left(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in H}\right)$
- Assume the rewards are deterministic, bounded in [0,1].
- we want to solve the problem :

$$\pi^* = \arg \max_{\pi \in \Pi_{\text{Markov,det}}} \rho(R^\pi)$$

  Where ρ is functional called risk measure

In practice, the risk measure satisfies certain natural properties from a measure of risk like monotonicity and translation invariance.

Risk measures

- **Threshold probability**

$$\Pr\left(R^{\pi} \geq T\right)$$

- **Value-at-Risk** At level $\alpha \in (0, 1)$:

$$\mathrm{VaR}_{\alpha}\left[R^{\pi}\right] = \inf\left\{x \in \mathbb{R} : \Pr\left(R^{\pi} \leq x\right) \geq \alpha\right\}.$$

- **Conditional-Value-at-Risk** At level $\alpha \in (0, 1)$:

$$\mathrm{CVaR}_{\alpha}\left[R^{\pi}\right] = \frac{1}{\alpha}\int_{0}^{\alpha} \mathrm{VaR}_{\gamma}\left[R^{\pi}\right] d\gamma$$

Entropic risk measure

For a random variable X and parameter $\beta \in \mathbb{R}$, the entropic risk measure is defined as

$$\rho_\beta(X) = \begin{cases} \frac{1}{\beta} \log \mathbb{E}\left[e^{\beta X}\right] & \beta \neq 0 \\ \mathbb{E}[X] & \beta = 0 \end{cases}$$

- Second order expansion of the entropic risk measure

$$\rho_\beta(X) = \mathbb{E}[X] + \frac{\beta}{2} \mathrm{Var}(X) + O(\beta^2), \qquad \beta \to 0.$$

- Can be exactly optimized using dynamic programming [4]
- Apart from the expectation, it's the only continuous objective that can be optimized exactly using dynamic programming[10]

Bellman equations for entropic risk measure

- **Planning in MDPs:** Entropic risk measure admit a *log-sum-exp* Bellman update with policy/value iteration for fixed risk parameter [7]: **Policy evaluation (reward, sums):**

$$V_t^\pi(s) = r_t(s,a) + \frac{1}{\beta} \log \left(\sum_{a \in \mathcal{A}(s)} \pi_t(a \mid s) \sum_{s' \in \mathcal{S}} P_t(s' \mid s, a) \exp\left\{ \beta\left(V_{t+1}^\pi(s')\right) \right\} \right)$$

  Optimality (reward, sums):

$$V_t(s) = r_t(s,a) + \max_{a \in A} \frac{1}{\beta} \log \left(\sum_{s' \in \mathcal{S}} P_t(s' \mid s, a) \exp\left\{ \beta\left(V_{t+1}(s')\right) \right\} \right)$$

- **Learning (RL):** Asymptotic convergence for risk-sensitive actor–critic and Q-learning [3, 2]

Best policy identification

We consider a BPI algorithm with a policy sequence $\{\pi^t\}_{t\in\mathbb{N}}$, an exploration budget τ, and an output policy $\hat{\pi}$.
Our goal is to design an (ε, δ)-PAC algorithm (defined below) that minimizes the sample complexity, i.e., the number of exploration episodes τ.

Definition (PAC algorithm for BPI)

An algorithm is (ε, δ)-PAC for best policy identification if it returns a policy $\hat{\pi}$ after some number of episodes τ that satisfies

$$\mathbb{P}\left(V_1^\star(s_1) - V_1^{\hat{\pi}}(s_1) \leq \varepsilon\right) \geq 1 - \delta$$

BPI problem challenges

The three core challenges are:

1. the exploration rule: How to explore the environment
2. the stopping rule: when to stop so that the output is (ε, δ)-PAC while minimizing the number of exploration episodes τ.
3. which policy to output

Contributions

- Literature review and problem framing
- Lower-bound on the BPI for entropic risk measure
- KL-driven exploration for entropic risk

Outline

Interaction with the Environment

- **Generative model (simulator):** Can simulate the next step from any (s, a)
- **Forward / dynamics model:** Can only interact with the environment. After sampling a state s_0, we sample a trajectory
- **Reward-free model:** Can only sample trajectories and we do not have access to the reward

Literature review for risk-neutral case

Perspective	Reference	Lower Bound	Upper Bound
Generative model	[1]	$\Omega\left(\frac{SAH^3}{\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Forward model	[5]	$\Omega\left(\frac{SAH^3}{\varepsilon^2}\right)$	
	[9]		$\tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\right)$
	[11]		$\tilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\right)$
Reward-free	[5]	$\Omega\left(\frac{SAH^3}{\varepsilon^2}+S\right)$	
	[8]		$\tilde{\mathcal{O}}\left(\frac{S^2AH^7}{\varepsilon}+\frac{S^2AH^5}{\varepsilon^2}\right)$
	[9]		$\tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}+S\right)$
	[11]		$\tilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}+S\right)$

Literature review for risk-sensitive case

Perspective	Reference	Lower Bound	Sampling complexity
Generative Model	[12]	$\Omega\left(\frac{SA\gamma^2}{c_1\varepsilon^2}\frac{e^{\lvert\beta\rvert\frac{1}{1-\gamma}}-3}{\lvert\beta\rvert^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{SA\left(S+\log(SA/\delta)\right)}{\varepsilon^2(1-\gamma)^2\beta^2}\cdot e^{2\lvert\beta\rvert\frac{1}{1-\gamma}}\right)$

Gap to lower bounds: Between the lower bound of [12] and current best upper bounds, there remains at least a gap of $H^2\left(e^{\lvert\beta\rvert H}-1\right)$

Open problems

- **Planning vs. learning:** Planning for entropic–risk MDPs is well understood, but *learning* remains less developed and optimal PAC/BPI guarantees are still open.
- **Tail amplification & variance:** The entropic criterion applies an exponential tilt $e^{\beta R}$ that magnifies tail outcomes—inflating high rewards when $\beta > 0$ (penalizing low rewards when $\beta < 0$)—which raises estimator variance and induces difficulty that scales exponentially in $|\beta|$ and the horizon H.
- **Regret results insufficient:** Even the strongest regret bounds to date do not close this gap; PAC guarantees tailored to entropic risk remain scarce.

Outline

Lower bound from the hard MDP

Theorem (PAC lower bound for the entropic value objective)

Fix $\beta \in \mathbb{R}$ and integers $H \geq 3$, $S \geq 4$, $A \geq 2$, $\delta \in (0, \frac{S-3}{2})$, and sufficiently small $\varepsilon > 0$. There exists an episodic MDP with horizon H, S states, and A actions such that any (ε, δ)-PAC algorithm must, for some instance of this MDP, use an expected number of episodes T satisfying

$$T = \Omega\Big(SAH^2 \frac{e^{|\beta| H} - 1}{\beta^2 \varepsilon^2} \log\left(\frac{S}{\delta}\right)\Big)$$

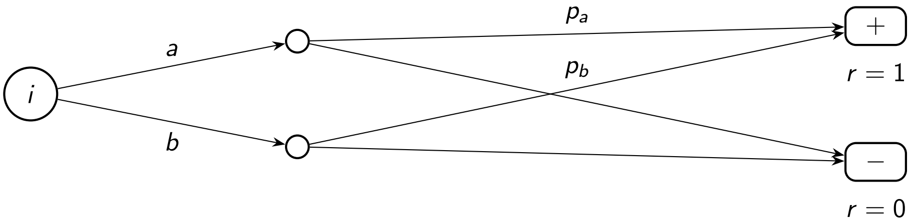
LeCam's method

Theorem

Fix θ_0,$\theta_1 \in \Theta$ *such that* $\rho(\theta_0, \theta_1) \geq 2\varepsilon$, *then :*

$$\inf_{\widehat{\theta}} \sup_{i \in \{0,1\}} P_{\theta}^{\otimes n}(\rho(\widehat{\theta}, \theta) \geq \varepsilon) \geq \frac{1 - \|P_{\theta_0}^{\otimes n} - P_{\theta_1}^{\otimes n}\|_{TV}}{2}$$

A hard bandit for a lower bound



Bandit Entropic Lower Bound

- **Create two hard instances:** they differ only in which action is better, and the advantage is tuned to be very small (exact target gap)
- **Bound information per episode:** the Kullback–Leibler divergence between the two instances for one episode is small so the two instances are hard to distinguish
- **Link indistinguishability to error** Apply the Bretagnolle–Huber inequality to show that if the two instances are hard to tell apart, any algorithm must make an error on at least one of them.
- **Trade accuracy for samples:** requiring small error on both instances forces a minimum total information budget.

Create two hard instances

We define two MDPs $\mathcal{M}^+$ and $\mathcal{M}^-$ within the family of MDPs given before that only differ in which Left action is better a or b.
Denote

$$p = \frac{1}{2c} \text{ and } q = p(1+\eta) \text{ and } c = e^{\beta R} - 1$$

Where η will be chosen to make the entropic gap between the two MDPs exactly ε, more precisely :

$$G(p) - G(q) = \frac{1}{\beta} \log\left(\frac{1+cp}{1+cq}\right) = \frac{1}{\beta}\log(1+\frac{\eta}{3})$$

Hence, we chose

$$\eta = 3(e^{\beta\varepsilon} - 1) \quad \text{So that} \quad G(p) - G(q) \leq \varepsilon$$

In $\mathcal{M}^+$, $p_a = q \quad$ and $\quad p_b = p$

Bound information per episode

Let $\mathbb{P}^+_{1:n}$ and $\mathbb{P}^-_{1:n}$ denote the laws of the entire n-episode simulation under $\mathcal{M}^+$ and $\mathcal{M}^-$. By decomposing the KL divergence we get:

$$\mathrm{KL}\left(\mathbb{P}^+_{1:n} \| \mathbb{P}^-_{1:n}\right) = \sum_{t=1}^{n} \mathbb{E}\big[\mathrm{KL}(\text{episode } t \mid \text{history})\big] \leq \; n \max\{d(p,q), d(q,p)\} \leq n \frac{\eta^2}{2c-1}$$

Link in-distinguishability to error

We apply the Bretagnolle–Huber inequality to the event A "the algorithm outputs action a". On $\mathcal{M}^+$ the error is A^{c}; on $\mathcal{M}^-$ the error is A :

$$\Pr_{\mathcal{M}^+}(A^{\mathrm{c}}) + \Pr_{\mathcal{M}^-}(A) \;\geq\; \tfrac{1}{2}\exp\Big(-\mathrm{KL}\left(\mathbb{P}^+_{1:n}\middle\|\mathbb{P}^-_{1:n}\right)\Big)$$

If the learner is (ε, δ)-correct on both instances, the LHS $\leq 2\delta$. Hence :

$$2\delta \;\geq\; \frac{1}{2}\exp\left(-n\cdot\frac{\eta^2}{2c-1}\right),$$

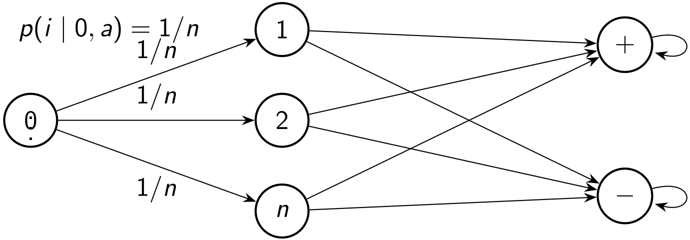
Lower bound for the hard bandit

Theorem

Any algorithm that solves the (ε, δ)*-PAC problem for* ε *satisfying* $\beta\varepsilon \leq \ln(2)$ *must have a sampling complexity that satisfy :*

$$n \;\geq\; \frac{2(e^{\beta R}-1)-1}{72\beta^2\varepsilon^2}\log\frac{1}{4\delta}$$

A hard MDP



$r(+) = H - 2$
$p(+ \mid i, a) = p_+ \quad \text{if } a = a_i^*$
$p(+ \mid i, a) = p_- \quad \text{if } a \neq a_i^*$

Lower bound from the hard MDP

Theorem (PAC lower bound for the entropic value objective)

Fix $\beta \in \mathbb{R}$ and integers $H \geq 3$, $S \geq 4$, $A \geq 2$, $\delta \in (0, \frac{S-3}{2})$, and sufficiently small $\varepsilon > 0$. There exists an episodic MDP with horizon H, S states, and A actions such that any (ε, δ)-PAC algorithm must, for some instance of this MDP, use an expected number of episodes T satisfying

$$T = \Omega\Big(SAH^2 \frac{e^{|\beta|H} - 1}{\beta^2 \varepsilon^2} \log\Big(\frac{S}{\delta}\Big)\Big)$$

Outline

Empirical MDP

Let $(s_h^i, a_h^i, s_{h+1}^i)$ be the state, the action, and the next state observed by an algorithm at step h of episode i.

For any step $h \in [H]$ and any state–action pair $(s, a) \in \mathcal{S} \times \mathcal{A}$, define

$$n_h^t(s, a) \triangleq \sum_{i=1}^{t} \mathbf{1}\{(s_h^i, a_h^i) = (s, a)\}, \qquad n_h^t(s, a, s') \triangleq \sum_{i=1}^{t} \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s, a, s')\}.$$

These definitions permit us to define the empirical transitions:

$$\hat{p}_h^t(s' \mid s, a) \triangleq \begin{cases} \dfrac{n_h^t(s, a, s')}{n_h^t(s, a)}, & \text{if } n_h^t(s, a) > 0, \\ \dfrac{1}{S}, & \text{otherwise.} \end{cases}$$

UCB algorithm

Sampling rule. At iteration t, act greedily w.r.t. $\widetilde{Q}^{\,t}$:

$$\forall s \in \mathcal{S},\ \forall h \in [H], \qquad \pi_h^{t+1}(s) \;=\; \arg\max_{a\in\mathcal{A}}\ \widetilde{Q}_h^t(s,a).$$

Stopping rule. Stop at the first t such that the certificate is below ε:

$$\tau \;=\; \inf\Big\{\, t \in \mathbb{N} :\ \pi_1^{t+1} G_1^{\,t}(s_1) \;\le\; \varepsilon \,\Big\}.$$

Prediction rule. Output the policy from the next iteration:

$$\widehat{\pi} \;=\; \pi^{\tau+1}.$$

Optimism

Given a confidence region set $\mathcal{U}_h^t(s,a)$ such that for any (s,a):

$$p_h(.|s,a) \in \mathcal{U}_h^t(s,a)$$

An optimistic upper bound of the value function is then:

$$Q_h(s,a) = r_h(s,a) + \sup_{p \in \mathcal{U}_h^t(s,a)} \rho_\beta(V_h^t(s,a))$$

$$V_h^t(s,a) = \max_{a \in A} Q_h^t(s,a)$$

We chose the KL-divergence confidence region:

$$\mathcal{U}_h^t(s,a) = \left\{ q \in \Sigma_S | D_{KL}(\widehat{p}_h^t(s,a) || q(s,a)) \leq \frac{\lambda\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)} \right\}$$

Why KL planning ?

1. For $\lambda(n,\delta) = \log\left(\frac{3SAH}{\delta}\right) + S\log\big(8e(n+1)\big)$ with probability $1-\delta$[11]:

$$p_h(s,a) \in \mathcal{U}_h^t(s,a)$$

2. Controls trajectory mismatch. The chain rule of KL yields :

$$\mathrm{KL}\big(\hat{P}^{t,\pi}, P^{\pi}\big) = \sum_{h,s,a} d_h^{t,\pi}(s,a)\,\mathrm{KL}\big(\hat{p}_h^t(\cdot|s,a),\ p_h^{\star}(\cdot|s,a)\big)$$

so shrinking local KL terms reduces global error.

A Bernstein's inequality

To turn the optimistic upper bound to a computable bound we use Bernstein's inequality [11]:

Theorem

For any $q \in \mathcal{U}_h^t(s,a)$*, for any bounded* $f : \mathcal{S} \to [0,b]$*:*

$$\left| E_p[f] - E_q[f] \right| \leq \sqrt{2\,\mathrm{Var}_q(f)\frac{\lambda(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \frac{2}{3}\,b\,\frac{\lambda(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$

Bonus terms

Using Bernstein's inequality, we define bonus terms :

$$\widetilde{B}_h^t(s,a) \triangleq 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}\big(e^{\beta \widetilde{V}_{h+1}^t}\big)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + e^{\beta H}(8H + 2\sqrt{2} + 3)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$

$$\widetilde{s}_h^t(s,a) \triangleq \frac{1}{\max\big\{1,\ \widehat{p}_h^t \underline{V}_{h+1}^t(s,a) - \widetilde{B}_h^t(s,a)\big\}}$$

Optimistic/pessimistic state-value function

Upper and lower Q- and value functions:

$$\widetilde{Q}_h^t(s,a) \triangleq \min\Bigg\{H,\ r_h(s,a) + \frac{1}{|\beta|}\log\big(1 + \widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)\big)$$

$$+ \frac{1}{\beta H}\Big(\rho_\beta^{\widehat{p}_h^t}\big(\widetilde{V}_{h+1}^t(s,a)\big) - \rho_\beta^{\widehat{p}_h^t}\big(\underline{V}_{h+1}^t(s,a)\big)\Big) + \rho_\beta^{\widehat{p}_h^t}\big(\widetilde{V}_{h+1}^t(s,a)\big)\Bigg\},$$

$$\underline{Q}_h^t(s,a) \triangleq \max\Bigg\{0,\ r_h(s,a) - \frac{1}{|\beta|}\log\big(1 + \widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)\big)$$

$$- \Big(\rho_\beta^{\widehat{p}_h^t}\big(\widetilde{V}_{h+1}^t(s,a) - \rho_\beta^{\widehat{p}_h^t}\big(\underline{V}_{h+1}^t(s,a)\big)\Big) + \rho_\beta^{\widehat{p}_h^t}\big(\underline{V}_{h+1}^t(s,a)\big)\Bigg\},$$

$$\widetilde{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}} \widetilde{Q}_h^t(s,a), \qquad \underline{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}} \underline{Q}_h^t(s,a), \qquad \widetilde{V}_{H+1}^t = \underline{V}_{H+1}^t \equiv 0.$$

Stopping rule

We define the stopping rule bounding the gap between the optimal policy and the policy at instant t:

$$G_h^t(s) = \min\left\{H, \frac{2}{\beta}\log\left(1+\widetilde{s}_h^t\widetilde{B}_h^t\right)\left(s, \pi_h^{t+1}(s)\right)+\left(1+\frac{3}{H}\right)\widehat{p}_h^t\pi_{h+1}^{t+1}G_{h+1}^t(s,a)\right)\right\} \qquad G_{H+1}^t \equiv 0$$

Optimism results

Lemma (Optimism and pessimism, entropic case)

On $\mathcal{G}$, for all $t \geq 0$, $h \in [H]$, and (s,a),

$$\underline{Q}_h^t(s,a) \leq Q_h^\star(s,a) \leq \widetilde{Q}_h^t(s,a) \qquad \underline{V}_h^t(s) \;\leq\; V_h^\star(s) \leq \widetilde{V}_h^t(s)$$

And the stopping rule :

Lemma

The greedy policy satisfies the standard gap domination:

$$V_1^\star(s_1) - V_1^{\pi^{t+1}}(s_1) \leq \pi_1^{t+1} G_1^t(s_1) \qquad \forall t \geq 0$$

Algorithm 1 Entropic-BPI (greedy w.r.t. upper entropic confidence)

Input: β, $\delta \in (0,1)$, $\varepsilon > 0$.
Initialize counts $n_h^0(\cdot) = 0$ and $\widehat{p}_h^0(\cdot|s,a) = 1/S$.
for $t = 0, 1, 2, \ldots$ **do**
  For $h = 1, \ldots, H$ Compute $\widetilde{B}_h^t$ and $\widetilde{s}_h^t$; then $(\widetilde{Q}_h^t, \widetilde{V}_h^t)$ and $(\underline{Q}_h^t, \underline{V}_h^t)$.
  Sampling rule: $\pi_h^{t+1}(s) \in \arg\max_{a \in \mathcal{A}} \widetilde{Q}_h^t(s,a)$ for all h, s.
  Stopping potential: for $a = \pi_h^{t+1}(s)$, set

$$G_h^t(s) = \min\Big\{H, \frac{2}{\beta}\log\big(1 + \widetilde{s}_h^t \widetilde{B}_h^t\big)\big(s, \pi_h^{t+1}(s)\big) + \Big(1 + \frac{3}{H}\Big)\widehat{p}_h^t \pi_{h+1}^{t+1} G_{h+1}^t(s,a)\Big)\Big\}$$

  Stopping rule: $\tau = \inf\big\{t \in \mathbb{N} : \ \pi_1^{t+1} G_1^t(s_1) \le \varepsilon\big\}$. If $t \ge \tau$, **stop** and output π^{t+1}.
  Execute episode $t+1$ with π^{t+1}, update counts and $\widehat{p}_h^{t+1}$.
end for

Sampling complexity

Theorem

For $\varepsilon \in]0,1]$ and $\delta > 0$ the algorithm Entropic-KL-BPI return an ε-optimal policy with probability at least $1-\delta$ after τ steps, where with probability of $1-\delta$ and we have an upper bound on τ:

$$\tau = \tilde{O}\left(\frac{(e^{|\beta|H}-1)^2}{\beta^2} \cdot \frac{H^2 SA}{\varepsilon^2} \log\left(\frac{3SAH}{\delta}\right)\right)$$

Outline

A promising approach

Planning using Renyi divergence:

$$Q_h^t(s,a) = r_h(s,a) + \sup_{p \in \mathcal{U}_{h,\kappa}^t(s,a)} \rho_\beta^p\big(V_{h+1}^t\big) \qquad V_h^t(s) = \max_{a \in \mathcal{A}} Q_h^t(s,a), \quad V_{H+1}^t \equiv 0$$

$$\mathcal{U}_{h,\kappa}^t(s,a) := \Big\{p : \; R_\kappa\big(p \| \hat{p}_h^t(\cdot|s,a)\big) \le r_h^t(s,a)\Big\}$$

We get by using the variational formula

$$\sup_{p : R_\kappa(p\|\hat{p}) \le r} \rho_\beta^p(X) \le \rho_\gamma^{\hat{p}}(X) + \frac{r}{\gamma - \beta} \qquad \gamma > \beta, \quad \kappa = \frac{\gamma}{\gamma - \beta}$$

This gives a computable UCB bound:

$$Q_h^t(s,a) \le r_h(s,a) + \rho_\gamma^{\hat{p}_h^t(\cdot|s,a)}\big(V_{h+1}^t\big) + \frac{r_h^t(s,a)}{\gamma - \beta} \qquad \gamma > \beta \quad \kappa = \frac{\gamma}{\gamma - \beta}$$

Renyi divergence

An optimal upper bound to the state-value function is:

$$Q_h^t(s,a) = r_h(s,a) + \sup_{p \in \mathcal{U}_{h,\kappa}^t(s,a)} \rho_\beta^p\left(V_{h+1}^t\right) \qquad V_h^t(s) = \max_{a \in \mathcal{A}} Q_h^t(s,a), \quad V_{H+1}^t \equiv 0$$

$$\mathcal{U}_{h,\kappa}^t(s,a) := \left\{p : \; R_\kappa\left(p \| \hat{p}_h^t(\cdot|s,a)\right) \leq r_h^t(s,a)\right\}$$

And we get by using the variational fomula:

$$\sup_{p \in \mathcal{U}_{h,\kappa}^t(s,a)} \rho_\beta^p\left(V_{h+1}^t\right) \leq \rho_\gamma^{\hat{p}}(X) + \frac{r}{\gamma - \beta} \qquad \gamma > \beta, \quad \kappa = \frac{\gamma}{\gamma - \beta}$$

This gives a computable UCB bound whenever $p_h(s,a) \in \mathcal{U}_{h,k}^t(s,a)$:

$$Q_h^t(s,a) \leq r_h(s,a) + \rho_\gamma^{\hat{p}_h^t(\cdot|s,a)}\left(V_{h+1}^t\right) + \frac{r_h^t(s,a)}{\gamma - \beta} \qquad \gamma > \beta \quad \kappa = \frac{\gamma}{\gamma - \beta}$$

Problem

We have to control the Renyi divergence

Renyi divergence is hard to control

Theorem

Let $\alpha > 1$ and let $\hat{p}_n$ be the empirical distribution of n i.i.d. samples from an unknown categorical distribution p on a finite alphabet. There exist constants $C_\alpha > 0$ and $\delta_0 \in (0,1)$ such that for every $\delta \in (0, \delta_0]$ there exists a (binary) distribution p with

$$\mathbb{P}_p\left(\exists n \geq 1 : nD_\alpha\big(\hat{p}_n \| p\big) \geq C_\alpha \delta^{-(\alpha-1)}\right) \geq \delta.$$

Consequently, any distribution-free, time-uniform tail inequality of the form

$$\sup_p \mathbb{P}_p\left(\exists n \geq 1 : nD_\alpha\big(\hat{p}_n \| p\big) \geq b(n, \delta)\right) \leq \delta$$

must satisfy, for each $\delta \in (0, \delta_0]$, that $b(n, \delta) \geq C_\alpha \delta^{-(\alpha-1)}$ for some n. In particular, no schedule with $b(n, \delta) = O(\log(1/\delta))$ works

Possible solutions (a burn-in phase)

Theorem (All-time Rényi bound after Chernoff burn-in)

Let p be a probability distribution on S, let $\hat{p}_n$ be the empirical estimation, and suppose $b = \min_{i,p_i>0} p_i$. Fix $\delta \in (0,1)$ and $\alpha > 1$ Set the likelihood–ratio cap target to $L_0 = 1 + \eta$ with $\eta = \frac{1}{2}$ (so $L_0 = \frac{3}{2}$), and define

$$n_0 = \left\lceil \frac{12}{b} \left(\log \frac{4S}{\delta} + \log \frac{24}{b} \right) \right\rceil$$

For $\varepsilon \in (0,1)$ define

$$\beta(n, \delta) = \log \frac{1}{\delta} + (S-1) \log \left(e \left(1 + \tfrac{n}{S-1} \right) \right)$$

Then, with probability at least $1-\delta$, simultaneously for all $n \geq n_0$,

$$D_\alpha(\hat{p}_n \| p) \leq \begin{cases} \frac{3}{n} \beta\left(n, \frac{\delta}{2}\right) & 1 < \alpha \leq 2 \\ \frac{\alpha (3/2)^{\alpha-1}}{n} \beta\left(n, \frac{\delta}{2}\right) & \alpha > 2 \end{cases}$$

References

[1] Mohammad Gheshlaghi Azar, Rémi Munos, and Hilbert J. Kappen. On the sample complexity of reinforcement learning with a generative model. In *Proceedings of the 29th International Conference on Machine Learning (ICML)*, pages 1707–1714, 2012.

[2] Vivek S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 26(2): 294–311, 2001. doi: 10.1287/moor.26.2.294.10559.

[3] Vivek S. Borkar. A sensitivity formula for risk-sensitive cost and the actor–critic algorithm. *Systems & Control Letters*, 44(5):339–346, 2001. doi: 10.1016/S0167-6911(01)00105-4.

[4] Vivek S. Borkar. Q-learning for risk-sensitive control. *Mathematics of Operations Research*, 27(2): 294–311, 2002. doi: 10.1287/moor.27.2.294.3238.

[5] Christoph Dann and Emma Brunskill. Sample complexity of episodic fixed-horizon reinforcement learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 2818–2826, 2015.

[6] Farama Foundation. Cliff walking — gymnasium. URL `https://gymnasium.farama.org/environments/toy_text/cliff_walking/`.

[7] Ronald A. Howard and James E. Matheson. Risk-sensitive markov decision processes. *Management Science*, 18(7):356–369, 1972. doi: 10.1287/mnsc.18.7.356.

[8] Chi Jin, Akshay Krishnamurthy, Max Simchowitz, and Tiancheng Yu. Reward-free exploration for reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning (ICML)*, volume 119 of *Proceedings of Machine Learning Research*, pages 4870–4879. PMLR, 2020. URL `https://proceedings.mlr.press/v119/jin20d.html`.

[9] Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and