école normale supérieure paris—saclay

Tübingen AI Center

Research internship - MVA master

Risk-sensitive Reinforcement learning

Report

Amer Essakine
supervised by
Dr Claire Vernade

—-

Abstract

We study risk-sensitive reinforcement learning in episodic MDPs through the entropic risk measure—one of the few continuous objectives that admits exact dynamic programming—motivated by applications where tail events matter and risk-neutral criteria are inadequate. We (i) synthesize the literature on sampling-complexity for pure exploration in both risk-neutral and risk-sensitive settings, framing open questions around PAC guarantees under entropic risk; (ii) prove information-theoretic lower bounds for best-policy identification that reveal an intrinsic "exponential curse": any (ε, δ)-PAC learner must pay a cost that grows exponentially in the product $|\beta|H$ (risk parameter $\times$ horizon), showing such dependence is unavoidable and scales with the size of the state–action space; and (iii) develop a UCB-style algorithm, **Entropic-KL-BPI**, that plans inside KL confidence regions and uses variance-aware bonus terms. We establish optimism/pessimism bounds and a stopping rule yielding a PAC sample-complexity upper bound

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

Finally, we explore a promising "tilted-variance" route that reasons under the exponentially tilted law to avoid the blow-up from $\mathrm{Var}(e^{\beta V})$. We derive a Bennett-type bias–variance control in terms of Rényi divergence, and show a limitation: distribution-free, time-uniform concentration for Rényi divergence cannot hold with logarithmic dependence on $1/\delta$ without additional conditions—motivating a Chernoff burn-in remedy. Overall, the report clarifies fundamental hardness for entropic objectives, provides the first PAC analysis of a KL-driven exploration strategy tailored to entropic risk, and charts a path toward sharper rates via tilted-measure analysis.

Abstract

[Résumé] Nous étudions l'apprentissage par renforcement sensible au risque dans les MDPs épisodiques au travers de la mesure de risque entropique — l'un des rares objectifs continus qui admet une programmation dynamique exacte — motivés par des applications où les événements rares importent et où les critères neutres au risque s'avèrent insuffisants. Nous (i) synthétisons la littérature sur la complexité d'échantillonnage pour l'exploration pure, tant en contexte neutre qu'en contexte sensible au risque, en formulant des questions ouvertes autour des garanties PAC sous risque entropique ; (ii) démontrons des bornes inférieures d'ordre informationnel pour l'identification de la meilleure politique qui révèlent une « malédiction exponentielle » intrinsèque : tout apprenant (ε, δ)-PAC doit payer un coût qui croît exponentiellement avec le produit $|\beta|H$ (paramètre de risque $\times$ horizon), montrant que cette dépendance est inévitable et s'accroît avec la taille de l'espace état–action ; et (iii) développons un algorithme de type UCB, **Entropic-KL-BPI**, qui planifie à l'intérieur de régions de confiance KL et utilise des termes bonus sensibles à la variance. Nous établissons des bornes d'optimisme/pessimisme et une règle d'arrêt donnant une borne supérieure PAC de la complexité d'échantillonnage :

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

Enfin, nous explorons une piste prometteuse dite de la « variance inclinée » qui raisonne sous la loi exponentiellement inclinée afin d'éviter l'explosion de $\mathrm{Var}(e^{\beta V})$. Nous dérivons un contrôle biais–variance de type Bennett en termes de divergence de Rényi, et montrons une limitation : une concentration sans hypothèse, uniforme en temps, pour la divergence de Rényi ne peut tenir avec une dépendance logarithmique en $1/\delta$ sans conditions supplémentaires — ce qui motive un remède par préchauffage de type Chernoff. Dans l'ensemble, ce rapport clarifie la difficulté fondamentale des objectifs entropiques, fournit la première analyse PAC d'une stratégie d'exploration pilotée par la divergence KL adaptée au risque entropique, et trace une voie vers des taux plus précis via l'analyse par mesure inclinée.

Contents

1 Introduction

Traditional RL algorithms rest on maximizing expected discounted return, deeming an action optimal if it minimizes the expected discounted cost cost assuming future actions are also chosen optimally. This yields risk-neutral policies that can overlook low-probability events that have high impact. In many domains the problem must explicitly account for risk. Consider a standard investment setting: one cares not only about expected profit but also about variability. Given two investments with nearly equal expected profit but very different variance, most investors prefer the lower-variance option, which shows that expected value alone may not capture a decision maker's preferences. Risk-sensitive RL addresses this gap by evaluating returns through risk measures rather than raw expectations to quantify uncertainty in random costs, providing robustness to tail risks and greater flexibility, since the agent can choose a measure aligned with its goals and risk tolerance.

To provide some motivation for incorporating risk into decision making, we briefly describe two illustrative examples. First consider the St. Petersburg paradox. An agent is offered participation in a lottery for a fixed monetary cost. The lottery's outcome is determined by a sequence of fair coin tosses, terminating upon the first observation of tails. The agent receives a payout of 2^K where k is the number of times heads have come up, and the event has a probability 2^{k+1}. The agent must decide whether to accept or reject participation in the game. The expected payoff of this game diverges to infinity, so a criterion based on maximizing expected profit would suggest paying any price to enter. In practice people offer only a modest amount, which reflects aversion to extreme variability and the heavy-tailed nature of the payoff distribution, indicating the formulation of the problem must take into account such risk measures.

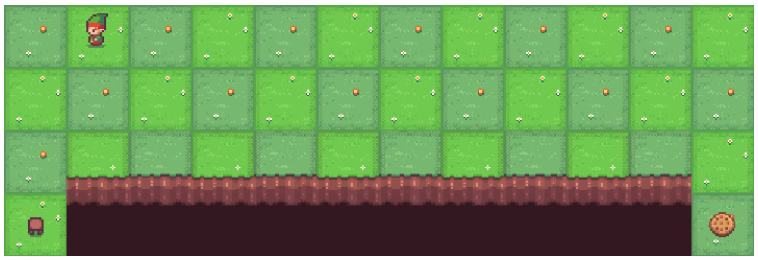

Figure 1: Cliff walking environment, a risky agent will always take the shortest path while a weary agent will take the long safe path

A second example is given by the Cliff walking problem. Consider a grid that has two zones : the feasible region and a cliff. The agent must travel from a start state to a goal across the grid. Each move incurs a small cost, falling off the cliff incurs a large terminal cost. Under a risk-neutral expected cost criterion the shortest path that hugs the edge can appear optimal, yet even a small slip or exploration noise creates a meaningful chance of catastrophe, so the

outcome distribution has a heavy lower tail. A risk-seeking objective prefers this risky route because it accepts higher variance to improve expected performance, whereas a risk-averse objective selects a slightly longer detour that sharply reduces failure probability. Risk-sensitive RL makes this explicit by optimizing a chosen risk measure. Tuning the measure and its parameters lets you move smoothly between risk-seeking and risk-averse behavior to match the application's goals and tolerances.

1.1 Markov decision process

In reinforcement learning, we have an agent that interacts with a dynamic environment by taking actions and receiving rewards at each step. The aim is to learn a strategy (a policy) that maximizes the cumulative rewards. Unlike supervised or unsupervised learning, where we are given a dataset and performance is evaluated on it, in reinforcement learning the data are generated by the agent's own behavior through its interactions with the environment. Using this feedback, the agent is *reinforced* to interact with the environment in the most optimal way and to take actions that maximize cumulative rewards over time.

More formally, we are given a set of states $\mathcal{S}$ describing the environment, a set of actions $\mathcal{A}$ available to the agent, and a horizon H representing the number of steps in each episode. At each step $h \in [H]$, the agent is in a state $s \in \mathcal{S}$ and takes an action $a \in \mathcal{A}$ according to its policy. The environment then provides a (possibly random) reward R_h with mean $r_h(s, a)$ and transitions to another state s'. In a stochastic environment, the next state is random and drawn according to the transition probability $p(s'|s, a)$.

The environment is typically assumed to be a Markov decision process (MDP), which satisfies the Markov property: conditional on the current state and action, the distribution of the next state (and reward) is independent of the past. More formally, for any $h \in [H]$, any sequence of states $s_{1:h+1} = (s_1, ..., s_h, s_{h+1}) \in \mathcal{S}$, and any sequence of actions $a_{1:h} = (a_1, ..., a_h) \in \mathcal{A}$, we have:

$$\Pr(S_{h+1} = s_{h+1} \mid S_{1:h} = s_{1:h}, A_{1:t} = a_{1:h}) = \Pr(S_{h+1} = s_{h+1} \mid S_h = s_h, A_h = a_h).$$

Equivalently, there exists a transition kernel $p_h(\cdot \mid s, a)$ such that $S_{h+1} \sim p_h(\cdot \mid S_h, A_h)$. We also assume a Markov reward model: For any measurable set B we have :

$$\mathbb{P}\big(R_h \in B \mid S_{1:h} = s_{1:h}, A_{1:h} = a_{1:h}\big) = \mathbb{P}\big(R_h \in B \mid S_h = s_h, A_h = a_h\big)$$

An MDP is then defined by the tuple

$$(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{R_h\}_{h\in[H]}),$$

where $\mathcal{S}$ is the set of states, $\mathcal{A}$ is the set of actions, H is the horizon (i.e., the number of steps in each episode), $p_h(s'|s, a)$ is the probability of transitioning to state s' from state s

after taking action a at step h, and $r_h(s,a)$ is the expected reward received.

The agent selects a strategy to follow while interacting with the environment which is represented by a collection of policies $\pi = ((\pi_h(.|s))_{s\in\mathcal{S}})_{h\in[H]}$ where $\pi_h(.|s)$ is a probability distribution over the action space given that the current state of the environment is s at the time step h. The goal is to learn a policy that maximizes the cumulative reward received by the agent while interacting with the environment.

For a policy π, we define the value function V^π which represents the expected reward (also called the return) accumulated from step h until the end of the horizon starting from state s :

$$V_h^\pi(s) \triangleq \mathbb{E}\left[\sum_{h=h'}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s\right] \quad \text{where } a_{h'} \sim \pi_{h'}(\cdot|s_{h'}) \text{ and } s_{h'+1} \sim p_{h'}(\cdot|s_{h'}, a_{h'})$$

The problem of reinforcement learning can then be formulated as an optimization problem where the initial state s_1:

$$\pi^\star \in \underset{\pi\in\Pi_{\text{Markov,stoch}}}{\operatorname{argmax}} \; V_1^\pi(s_1)$$

Where $\Pi_{\text{Markov,stoch}} = \left\{\pi = (\pi_h)_{h=1}^H : \pi_h : \mathcal{S} \to \Delta(\mathcal{A})\right\}$ is the set of stochastic Markovian policies, we denote $V_h^*(s) = \max_\pi V_h^\pi(s)$.

Notice that the optimization problem is equivalent to the problem when the initial state is not fixed but drawn from some distribution μ as we can always add an artificial initial state s_0 such that for any action a the transition probability is given by : $p(.|s_0, a) = \mu$ so we will always restrict to the case when the initial state is fixed.
It is well known that in finite-horizon MDPs it suffices to search for a deterministic, Markov (time-dependent) policy (Puterman, 1994, Theorem 4.4.2). A deterministic policy can be considered as a collection of functions $\pi_h : \mathcal{S} \to \mathcal{A}$ and the value function for a this class of policies is defined as

$$V_h^\pi(s) = \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'}, a_{h'}) \middle| s_h = s\right] \quad \text{where } a_{h'} = \pi_{h'}(s_{h'}) \text{ and } s_{h'+1} \sim p_{h'}(\cdot|s_{h'}, a_{h'})$$

And the Reinforcement learning problem is :

$$\pi^\star \in \underset{\pi\in\Pi_{\text{Markov,det}}}{\operatorname{argmax}} \; V_1^\pi(s_1)$$

Where $\Pi_{\text{Markov,det}} = \left\{\pi = (\pi_h)_{h=1}^H : \pi_h : \mathcal{S} \to \mathcal{A}\right\}$ is the set of deterministic Markovien policies.

We also define the state-action value function which represents the expected reward accumulated from step h until the end of the horizon starting from state s and taking action a

$$Q_h^\pi(s,a) \triangleq \mathbb{E}\left[\sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'})\middle| s_h = s \middle| a_h = a\right] \quad \text{where } a_{h'} = \pi_{h'}(s_{h'}) \text{ and } s_{h'+1} \sim p_{h'}(\cdot|s_{h'},a_{h'})$$

Perhaps one of the most important result in RL is the Bellman expectation equations and Bellman optimality equations ; For a fixed policy, the Bellman expectation equation says that the value of a state at a given time step equals the immediate reward you expect to receive now plus the value you expect to obtain later, after the environment moves you to a next state. More formally :

$$V_h^\pi(s) = \pi_h Q_h^\pi(s), \qquad \text{and} \quad Q_h^\pi(s,a) = r_h(s,a) + p_h V_{h+1}^\pi(s,a)$$

Where we use the notation $p_h f(s,a) \triangleq \mathbb{E}_{s'\sim p_h(\cdot|s,a)}[f(s')]$ denotes the expectation operator with respect to the transition probabilities p_h, and $(\pi_h g)(s) \triangleq \pi_h g(s) \triangleq g\big(s,\pi_h(s)\big)$ denotes the composition with the policy π at step h
On the other hand, The Bellman optimality equation says that from an optimal state, taking the best available action now and then acting optimally after achieves the optimal value :

$$V_h^\star(s) = \max_a Q_h^\star(s,a), \qquad \text{and} \quad Q_h^\star(s,a) = r_h(s,a) + p_h V_{h+1}^\star(s,a).$$

These equations form the basis for the dynamic programming methods. the expectation equation are used for example to evaluate the value function for a policy what's called the policy evaluation while the optimality equation are used to find the optimal policy using policy iteration or value iteration.

We call the setting in which the model (i.e., the transition probabilities and rewards) is known **planning**. Planning is well understood and admits efficient algorithms, but it is often unrealistic: in real-world applications, the agent learns about the environment gradually while interacting with it. This is the **learning** setting, which is harder because we must estimate the model while simultaneously seeking an optimal policy. Balancing these two objectives is the exploration–exploitation trade-off: should we continue exploring to learn more about the model, or exploit what we already know to maximize reward?

1.2 Risk sensitive in MDPs

In risk-sensitive RL in a finite episodic MDP $(\mathcal{S},\mathcal{A},H,\{p_h\}_{h\in[H]},\{r_h\}_{h\in[H]})$ We consider deterministic policies π which is a collection of function $\pi_h : \mathcal{S} \to \mathcal{A}$. We denote the cumulative reward as a random variable

$$R_h^\pi(s,a) = \sum_{h'=h}^{H} r_{h'}(s_{h'},a_{h'}) \quad \text{where } a_{h'} = \pi_{h'}(s_{h'}) \text{ and } s_{h'+1} \sim p_{h'}(\cdot|s_{h'},a_{h'})$$

and $R_h^\pi(s) = R_h^\pi(s, \pi_h(s))$ and $R^\pi = R_1^\pi(s_0)$. In our setting, we have the more generalized problem :

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

Where ρ is functional called risk measure.

1.2.1 Risk measures

The motivation behind risk measures comes from the theory of mathematical finance. Maximizing only the mean of the return ignores variability and tail events (rare disasters on windfalls) which could be catastrophic in market analysis for example. We then define a functional defined on some space of random variables that represents the monetary value of the risk with the goal of maximizing by a trading off average performance against variability and bad tails. In the context of RL, we ranks policies by $\rho(R^\pi)$ rather than $V_1^\pi = \mathbb{E}[R^\pi]$ so we prefer a steady policy with slightly lower mean over a spiky policy with catastrophic lows for risk-averse setting while a risk-seeking agent will often pick a spiky policy with rare big wins over a steadier one, even at the cost of more catastrophic lows, provided the upside sufficiently boosts the chosen objective.
To account for all possible worst cases, a risk measure is usually defined as a point-wise supremum of a base risk measure over an uncertainty set which is the set of acceptable returns. For a risk measure to be coherent(Artzner et al., 1999), it needs to verify some properties :

(i) **Monotonicity:** if $X \leq Y$, then $\rho(X) \geq \rho(Y)$, $\forall X, Y \in L^p$. A point wise larger payoff Y is never riskier than X. The axiom enforces basic order consistency: worse outcomes must not be assigned a lower risk.

(ii) **Translation Invariance:** $\rho(X + c) = \rho(X) - c$, $\forall X \in L^p$, $\forall c \in \mathbb{R}$. Adding a sure cash amount c to a position reduces its risk one-for-one which makes the risk measure compatible with capital add-ons and safe assets.

(iii) **Sub additivity:** $\rho(X + Y) \leq \rho(X) + \rho(Y)$, $\forall X, Y \in L^p$ Combining positions should not create more risk than managing them separately.

(iv) **Positive homogeneity:** $\rho(\lambda X) = \lambda \rho(X)$, $\forall X \in L^p$, $\forall \lambda \geq 0$. Scaling the size of a position scales its risk proportionally—risk.

A risk measure that verify each of this properties is called a *coherent risk measure*, and it's called a convex risk measure if instead of properties (ii), (iii) it only verify a convexity property :

$$\rho(\lambda X + (1 - \lambda)Y) \leq \lambda \rho(X) + (1 - \lambda)\rho(Y) \forall X, Y \in L^p \forall \lambda \in [0, 1]$$

Föllmer and Schied (2002) proved that a map $\rho : L^p \to \mathbb{R} \cup \{\infty\}$ is a convex risk measure if and only if it can be represented as :

$$\rho(X) = \sup_{q \in \mathcal{Q}} \big\{ qX - \alpha_\rho(q) \big\} \qquad \forall X \in L^p$$

$$\text{where} \qquad \alpha_\rho(q) = \sup_{X \in L^p} \big\{ qX - \rho(X) \big\}.$$

Moreover, ρ is finite if and only if the supremum over $\mathcal{Q}$ in its dual presentation is attained. We cite some relevant examples of risk measures :

Value at Risk (VaR) VaR represents the quantile of the distribution. At level $\alpha \in (0, 1)$,

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

VaR_α is the performance threshold that is met with probability at least α(i.e., the α–quantile). It is widely used for its simple interpretation but unfortunately it is neither a coherent nor a convex risk measure as it fails the sub-additivity and the convex property.

Conditional Value at Risk (CVaR). The CVaR at level α is given by

$$\mathrm{CVaR}_\alpha\big[R^\pi\big] = \frac{1}{\alpha} \int_0^\alpha \mathrm{VaR}_\gamma\big[R^\pi\big] d\gamma.$$

$\mathrm{CVaR}_\alpha[R^\pi]$ is the *average of the worst α–fraction of rewards*—the mean of the lower α-tail of the reward distribution. Maximizing CVaR therefore lifts the tail floor, producing conservative policies that protect against very low-reward outcomes. It is a more adapted risk measure than VaR as it is a coherent (and therefore convex) risk measure. We can express the CVaR as an optimization problem

$$\mathrm{CVaR}_\alpha[R] = \sup_{t \in \mathbb{R}} \Big\{ t - \frac{1}{\alpha} \mathbb{E}\big[(t - R)_+\big] \Big\}, \qquad (x)_+ = \max\{x, 0\}. \tag{1}$$

Which was used in the literature to maximize CVaR using actor-critic approaches.

Threshold probability (simple). Pick a target level T and choose a policy that makes it least likely to fall below that target:

$$\Pr\big(R^\pi \leq T\big)$$

We minimize the chance of under performing the target level T. This is dual to VaR optimization: fixing T and minimizing the shortfall probability is analogous to fixing α and maximizing VaR_α.

Those three risk measures benefit from a very natural interpretation that makes them widely used to quantify risk. However, maximizing them in an MDP is untractable. Indeed, Marthe et al. (2023) proved that only the mean reward and the entropic risk measure maximization problem can be solved using dynamic programming. For other measure, the problem become very hard and we are not even sure if the optimal policy is Markovien. One of the

approaches to deal with this problem is to augment the state space in the MDP with a continuous variable that tracks the cumulated reward so far (with the intuition of making the policy Markovien by keeping track of the history) but this makes the problem expensively computable.

1.2.2 Entropic risk measure

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

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

Interpretation. $\rho_\beta(X)$ is derived from the exponential-utility of X: it is the unique c such that $\mathbb{E}\left[e^{\beta(X-c)}\right] = 1$ which means it is the sure amount you'd accept instead of the risky X showcasing its valor in quantifying the risk in $Xbyanumber$. It is the scaled log–moment generating function, so it summarizes both mean and dispersion.
As $\beta \to 0$, $\rho_\beta(X) \to \mathbb{E}[X]$, and for small $|\beta|$:

$$\rho_\beta(X) \approx \mathbb{E}[X] + \tfrac{\beta}{2} \operatorname{Var}(X) + \mathcal{O}(\beta^2)$$

showing a mean–variance tradeoff. $\beta < 0$ yields a risk-averse criterion (penalizing variability and low rewards),while $\beta > 0$ is risk-seeking (emphasizing high rewards). It is a convex risk measure but not a coherent one as it does not satisfy the sub-additivity property nor the positive homogeneity. Being a convex risk measure, it has a dual presentation which is the famous Donsker–Varadhan variational formula : For $X \in L^1$ and $\beta \in \mathbb{R}$,

$$\rho_\beta[X] = \frac{1}{\beta} \log(pe^{\beta X}) = \begin{cases} \sup\limits_{q \ll p} \left\{ qX - \frac{1}{\beta} D_{\mathrm{KL}}(q||p) \right\} & \beta > 0, \\ \inf\limits_{q \ll p} \left\{ qX + \frac{1}{|\beta|} D_{\mathrm{KL}}(q||p) \right\} & \beta < 0 \\ \mathbb{E}[X] & \beta = 0 \end{cases}$$

where $D_{\mathrm{KL}}(q||p) = \mathbb{E}_{\mathbb{P}}\left[\frac{dq}{dp} \log \frac{dq}{dp}\right]$ is the KL divergence. When finite, the optimizer is the exponentially tilted law

$$\frac{dp_\beta}{dp} = \frac{e^{\beta X}}{\mathbb{E}[e^{\beta X}]}.$$

One of the main properties that make us interested in the entropic risk measure is that apart of the mean of the rewards it is the only continuous objective that can be optimized using dynamic programming (Marthe et al., 2023). Moreover, most relevant risk measures can be approximated using the entropic risk measure using Chernoff bounds, for the threshold

probability for example :

$$\begin{aligned}\min_{\pi} \Pr\left(R^{\pi} \leq T\right) &\leq \min_{\pi} \min_{\beta<0} \mathbb{E}\left[e^{\beta(R^{\pi}-T)}\right] \\ &= \min_{\beta<0} \min_{\pi} \mathbb{E}\left[e^{\beta(R^{\pi}-T)}\right] \\ &= \min_{\beta<0} \mathbb{E}\left[e^{\beta(R^{\pi^{\star}_{\beta}}-T)}\right]\end{aligned}$$

Similarly, for the $CVaR$ and VaR. Ahmadi-Javid (2012) introduced a coherent risk measure called the entropic value at risk :

$$\mathrm{EVaR}_{\alpha}[X] = \sup_{\beta<0} \left\{ \rho_{\beta}(X) - \frac{1}{\beta}\log(\alpha) \right\}.$$

The $EVaR$ is a coherent risk measure and is proven to be a good approximation of VaR and CVaR :

$$\mathrm{VaR}_{\alpha}[X] \geq \mathrm{CVaR}_{\alpha}[X] \geq \mathrm{EVaR}_{\alpha}[X]$$

This provide a promising approach to solving RL with risk measures objective such as $CVaR$ and VaR : we write them as an optimization problem in the entropic risk measure, we then solve the RL problem for different β to find π^{*}_{β} the optimal policy for the objective ρ_{β} and return π^{*}_{β} for the β that solves the optimization problem. More formally :

$$\beta^{\star}_{\text{(C)VaR}} = \arg\sup_{\beta<0} \left\{ \rho_{\beta}\left[R^{\pi^{\star}_{\beta}}\right] - \tfrac{1}{\beta}\log(\alpha) \right\}, \qquad \beta^{\star}_{\text{TP}} = \arg\inf_{\beta<0} \left\{ e^{-\beta T}\mathbb{E}\left[e^{\beta R^{\pi^{\star}_{\beta}}}\right] \right\}$$

As promising as the approach seems, the problem is still difficult as we need to solve for a continuous interval of β. Marthe et al. (2025) provide an algorithm capable of solving this problem in the planning setting by leveraging the regularity of the entropic risk measure. However, the approach does not transfer immediately to the learning where we do not know the model (i.e the reward and transition probabilities) and as we will see after, even solving the problem for a fixed β with a finite time guarantees is hard.

1.3 Contributions

In this report, we make the following contributions:

- **Literature review and problem framing** In Section 2, We provide a thorough review of advancements in risk-sensitive RL and give a survey of existing results for the best policy identification problem for risk-neutral and risk-sensitive Reinforcement learning, We give detailed highlight open problems and pinpoint gaps in existing guarantees.
- **Lower-bound on the BPI for entropic risk measure** In Section 3, We develop an information-theoretic lower bound on the sample complexity of risk-sensitive RL,

clarifying its dependence on the planning horizon, the size of the state–action space, and the risk parameter showing that an exponential depenedency on the horizon and the risk parameter is unavoidable.

- **KL-driven exploration for entropic risk** In Section 4. We introduce a principled exploration strategy tailored to entropic risk measures that is based on planning inside an estimated MDP within a KL confidence region. We outline the algorithmic design and provide theoretical/empirical support.
- **A promising approach** In section 5, we highlight an approach that we have been following that is based on the tilted distribution. Cite the results proven this far and the problems with the approach

2 Litterature review of sampling complexity problems

2.1 Risk-sensitive reinforcement learning

Early work on risk-sensitive reinforcement learning traces back to risk-sensitive optimal control (Whittle, 1990), motivated by the idea that the agent evaluates rewards through a utility function ($U(R)$ rather than the raw reward R itself (as in expected-utility theory). A concave U encodes risk-averse behavior, a convex U risk-seeking behavior, and $U(R) = R$ recovers the risk-neutral case. Among utility-based criteria, the entropic risk measure(exponential utility) has attracted particular interest since it exhibits constant absolute risk aversion, hence a risk attitude independent of the wealth (Pratt, 1964; Arrow, 1971). In linear–Gaussian control this leads to the classical LEQG solution with Riccati-type recursions (Jacobson, 1973). In the context of MDPs Howard and Matheson (1972) formalized risk-sensitive MDP with exponential utility and derived dynamic-programming equations with a log-sum-exp Bellman update, with policy and value iteration for a constant risk parameter; The planning setting was well understood with efficient algorithms for evaluation and improvement. On the learning side, a foundational work by Borkar and Meyn (2000) established asymptotic convergence of risk-sensitive RL algorithms using standard stochastic approximation theory : Borkar (2001b) developped and actor-critic method for the risk sensitive average-cost, and Borkar (2001a) proposed a Q-learning scheme for the entropic risk measure using the Bellman equation. While these works provided convergence guarantees, finite-time sample-complexity bounds and large-scale empirical validation were sill not established.

For other risk measures, Ruszczyński (2010) generalized the Bellman equations for conditional time consistent risk measures. Time consistency means that if a policy is judged less risky than another policy at some future step, then it is also less risky at the current step. This property yields a tower property : the value at each step equals the risk (given today's information) of the immediate reward plus the next-step value. And we recover the Bellman recursion. More practical risk measures like CVaR and VaR are whoever time in-consistent, so they do not admit a standard Bellman decomposition and can produce preference rever-

sals across steps (Detlefsen and Scandolo, 2005). Furthermore, Marthe et al. (2023) show that, among continuous risk measures only the entropic risk measure admit exact dynamic-programming solutions. One approach in the literature to restore the Bellman recursion is to augment the MDP state space with an auxiliary tail parameter using the Rockafellar–Uryasev representation of CVaR, thereby obtaining a Bellman-type recursion on the extended space with non smooth plus-terms (Rockafellar and Uryasev, 2000). Following the same intuition, Ávila Pires et al. (2025) develop distributional dynamic programming over stock-augmented state space and establish conditions on the statistical functions that are Bellman optimizable, admitting distributional value/policy iteration and a deep-RL DQN implementation (with applications to CVaR and VaR).

There exists other approaches besides dynamic programming. For the entropic risk measure in the average-cost setting, Borkar (2001b) derives a policy-gradient formula and then derive a two–time-scale actor–critic algorithm, proving almost-sure convergence under standard ergodicity and step-size conditions. In a complementary line, Tamar et al. (2015) derive a gradient formula derived by for coherent risk measures leveraging the fact that they admit a convex dual representation, using this formula they devise an actor-critic scheme with value-function approximation. Recent works push these ideas to Deep RL, Xiao et al. (2024) derive analytical expressions for the policy gradient in the distributional RL setting for coherent risk measures. Then they propose CDPG, which approximates the return distribution with a categorical head. A complementary variational route leverages the control-as-inference link to build a risk-sensitive variation actor-critic algorithm (Granados et al., 2025).

2.2 Best policy identification

In Best policy identification problem, the agent interacts with the MDP as described in the setting section and observes the reward. In each episode t, the agent follows a policy π^t (the sampling rule) based only on the information collected up to and including episode $t-1$. At the end of each episode, the agent can decide to stop collecting data (we denote by τ its random stopping time) and outputs a guess $\hat{\pi}$ for the optimal policy.

A BPI algorithm is therefore made of a triple $\left((\pi^t)_{t\in\mathbb{N}}, \tau, \hat{\pi}\right)$. The goal is to build an (ε, δ)-PAC algorithm according to the following definition, for which the *sample complexity*, that is the number of exploration episodes τ, is as small as possible.

Definition 1 (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*

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

How the agent interacts with the environment matters. Some works assume access to a generative model—an oracle that, for any state–action pair, returns an independent sample of the next state and reward. While analytically convenient, this assumption is often unrealistic. A more practical setting is a forward model: the agent can only roll out trajectories

from an initial distribution, choosing actions but unable to reset to arbitrary states. An even more restrictive (but different) setting is reward-free exploration, where rewards are unknown or withheld during the exploration phase where we learn the transition model of the MDP in a given number of steps and then in the second phase we do planning on this learned model to compute a near-optimal policy for a given reward model.

2.3 Risk-neutral case

This problem have been studied extensively from the three perspectives in the risk-neutral case:

Perspective	**Reference**	**Lower Bound**	**Upper Bound**
Generative model	(Azar et al., 2012)	$\tilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\right)$	$\tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\right)$
Forward model	(Dann and Brunskill, 2015)	$\tilde{\mathcal{O}}\left(\frac{SAH^2}{\varepsilon^2}\right)$	
	(Kaufmann et al., 2021)		$\tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}\right)$
	(Ménard et al., 2021)		$\tilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\right)$
Reward-free	(Dann and Brunskill, 2015)	$\tilde{\mathcal{O}}\left(\frac{SAH^2}{\varepsilon^2}+S\right)$	
	(Jin et al., 2020)		$\tilde{\mathcal{O}}\left(\frac{S^2AH^7}{\varepsilon}+\frac{S^2AH^5}{\varepsilon^2}\right)$
	(Kaufmann et al., 2021)		$\tilde{\mathcal{O}}\left(\frac{SAH^4}{\varepsilon^2}+S\right)$
	(Ménard et al., 2021)		$\tilde{\mathcal{O}}\left(\frac{SAH^3}{\varepsilon^2}\right)$

Generative model In the discounted infinite-horizon setting with access to a generative model Azar et al. (2012) provided a lower bound on the BPI problem: Any algorithm that outputs an $\varepsilon-$optimal policy with probability at least $1-\delta$ must make at least $\mathcal{O}\left(\frac{SA}{(1-\gamma)^3\varepsilon^2}\log\frac{SA}{\delta}\right)$ call to the oracle. This translate to the episodic MDP setting into a lower bound of $\mathcal{O}\left(\frac{SAH^3}{\varepsilon^2}\log\frac{SA}{\delta}\right)$. They also provide a direct model-based algorithm based on the Q-learning that achieves the upper bound $\mathcal{O}\left(\frac{SAH^4}{\varepsilon^2}\log\frac{SA}{\delta}\right)$ in the non-stationary case.

Forward model Dann and Brunskill (2015) established a lower bound in the stationary case on the number of trajectories needed to achieve an $\varepsilon-$optimal policy to be of the order of $\mathcal{O}\left(\frac{SAH^2}{\varepsilon^2}\log\frac{SA}{\delta}\right)$. There have been many gradual improvements for the risk neutral case, Kaufmann et al. (2021) proposed RF-UCLR, adapting the UCLR algorithm (Jaksch et al., 2010) to best policy identification. The algorithm consists of doing planning on a learned transistion models that exists within a confidence region of the true model, in RF-UCLR

we use ℓ^1-balls. In practice, we establish bonus term in this confidence region to guide exploration. This algorithm achieved a sampling complexity of $\mathcal{O}\left(\frac{SAH^4}{\varepsilon^2}\log\frac{SA}{\delta}\right)$ matching (Azar et al., 2012) upper bound but on a much harder problem. Ménard et al. (2021) used the same approach to derive BPI-UCBVI algorithm using KL divergence confidence regions instead of the ℓ^1 confidence region, this achieved a sampling complexity of $\mathcal{O}\left(\frac{SAH^3}{\varepsilon^2}\log\frac{SA}{\delta}\right)$ in the non-stationary case.

Reward Free framework Based on (Dann and Brunskill, 2015), Ménard et al. (2021) declare a lower bound on RF setting for the stationary case. Any algorithm that achieve a ε-optimal policy needs at least $\mathcal{O}\left(\frac{SAH^2}{\varepsilon^2}\left(\log(\frac{SA}{\delta})+S\right)\right)$ trajectories. Jin et al. (2020) propose RF-RL-explore algorithm, the exploration bonus is based on the EULER algorithm (Zanette and Brunskill, 2019) a UCB-type algorithm. For each (s,h) assigns a reward 1 when you arrive at state s at step h and 0 otherwise and run the Euler algorithm then make the action at (s,h) uniform so all actions there get coverage. Collect the resulting policies into a set and sample policies uniformly from this set to build a dataset, ensuring that any significant (s,a,h) is visited with probability proportional to its max-reachable probability. This provide a sampling complexity of $\mathcal{O}(\frac{H^7S^2A}{\varepsilon}\log^3(\frac{1}{\delta})+\frac{H^5S^2A}{\varepsilon^2}\log(\frac{1}{\delta}))$. Kaufmann et al. (2021) also provided an algorithm for free reward exploration based similarly on learning a transition model inside a confidence region defined by ℓ^1-distance, while Ménard et al. (2021) relied on KL divergence confidence regions, the papers achieved a sampling complexity of $\mathcal{O}\left(\frac{SAH^4}{\varepsilon^2}(\log(\frac{SA}{\delta})+S)\right)$ and $\mathcal{O}\left(\frac{SAH^3}{\varepsilon^2}\log\frac{SA}{\delta}\right)$ respectively.

2.4 Entropic risk measure

For the entropic risk measure, the problem is harder because it magnifies tail outcomes—inflating high rewards when $\beta > 0$ and penalizing low rewards when $\beta < 0$. Moreover, as this line of work is relatively recent and most results aim to establish regret bounds, results on PAC bounds remain scarce; to our knowledge, there are no reward-free algorithms with theoretical guarantees specifically tailored to this criterion.

Perspective	Reference	Lower Bound	Upper / Regret Bound
GM	(Mortensen and Talebi, 2025)	$\mathcal{O}\left(\frac{SA\gamma^2}{c_1\varepsilon^2}\frac{e^{\lvert\beta\rvert\frac{1}{1-\gamma}}-3}{\lvert\beta\rvert^2}\right)$	$\mathcal{O}\left(\frac{SA\gamma^2}{c_1\varepsilon^2}\frac{\left(e^{\lvert\beta\rvert\frac{1}{1-\gamma}}-1\right)^2}{\lvert\beta\rvert^2}\right)$

Generative models: In the discounted infinite-horizon **?** proved low bounds on the number of oracle calls needed for an $\varepsilon-$optimal policy. In particular, they showed than an exponential dependency on the risk parameter β and on the horizon H is unavoidable: Any algorithm that outputs an $\varepsilon-$optimal policy with probability at least $1-\delta$ must make at least

Perspective	Reference	Lower Bound	Upper / Regret Bound
Forward model	(Fei et al., 2020)		$\tilde{\mathcal{O}}\big(\lambda(\lvert\beta\rvert H^2)\sqrt{H^3S^2AT}\big)$
	(Fei et al., 2020)		$\tilde{\mathcal{O}}\big(\lambda(\lvert\beta\rvert H^2)\sqrt{H^4SAT}\big)$
	(Fei et al., 2021)		$\tilde{\mathcal{O}}\big(\frac{e^{\lvert\beta\rvert H}}{\lvert\beta\rvert H}\sqrt{H^4S^2AK}\big)$

$\mathcal{O}\left(\frac{SA\gamma^2}{c_1\varepsilon^2}\frac{e^{|\beta|\frac{1}{1-\gamma}}-3}{|\beta|^2}\log\left(\frac{S}{c_2\delta}\right)\right)$ call to the oracle. They also provided a model-based straightforward algorithm $\mathcal{O}\left(\frac{SA\gamma^2}{c_1\varepsilon^2}\frac{\left(e^{2|\beta|\frac{1}{1-\gamma}}-1\right)^2}{|\beta|^2}\log\left(\frac{SA}{c_2\delta}\right)\right)$. This gives the first explicit sample-complexity characterization for entropic risk measure objectice with a generative model.

Forward model: In the episodic trajectory setting, tight PAC/BPI sample-complexity bounds for the entropic risk measure is still an open problem; the literature focuses on regret bounds. The first non-asymptotic results are model-free optimistic algorithms RSVI/RSQ of Fei et al. (2020), which establish $\mathcal{O}\big(\lambda(|\beta|H^2)\sqrt{H^3S^2AT}\big)$ (RSVI) and $\mathcal{O}\big(\lambda(|\beta|H^2)\sqrt{H^4SAT}\big)$ (RSQ) regret with $\lambda(u) = (e^{3u}-1)/u$. Fei et al. (2021) introduce the *exponential Bellman equation* and a doubly-decaying bonus, removing an extra $e^{|\beta|H^2}$ factor and yielding a regret of
$\mathcal{O}\Big(\frac{e^{|\beta|H}-1}{|\beta|H}\sqrt{H^4S^2AK\log^2\left(HSAK/\delta\right)}\Big)$
2

2.5 Open problems

As we have seen from the literature review. While planing is well understood since the start of the twentieth century learning is still little understood and optimal PAC guarantees are still to be found. This is mainly because the entropic risk measure is harder For the entropic risk criterion, the exponential tilting $e^{\beta R}$ magnifies tail outcomes—inflating high rewards when $\beta > 0$ (or high costs in cost form with $(\beta > 0)$—which increases estimator variance and induces hardness that scales exponentially with $|\beta|$ and the horizon H, making learning provably harder than in the risk-neutral case.

Between the lower bound established by (Mortensen and Talebi, 2025) and the current optimal upper bounds there still exists a gap of at least $H^2(e^{|\beta|H}-1)$ and even the best existing results on the regret does not close the gap.

In fact, Consider and MDP $(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{R_h\}_{h\in[H]})$, and an algorithm that after T episodes return a collection of policies $\{\pi_t\}_{t=1}^T$ and incur a regret upper bounded by $\mathcal{K}(T,\delta)$

then with probability $1-\delta$:

$$\frac{1}{T}\sum_{t=1}^{T}\big(V^{\star}-V^{\pi_t}\big)\leq\frac{\mathcal{K}(T,\delta)}{T}$$

To transform this to a sampling complexity bound, we draw $\tau\sim\text{Unif}\{1,...,T\}$ at the start of the episode and simulate π_τ for this episode. By the total law of expectation we have :

$$V^{*}-V^{\pi_\tau}=\frac{1}{T}\sum_{t=1}^{T}(V^{*}-V^{\pi_t})$$

To guarantee an $\varepsilon-$optimal policy, we must chose T verifying :

$$\frac{\mathcal{K}(T,\delta)}{T}\leq\varepsilon$$

An optimal regret bound is of the form $c\sqrt{T\log(\frac{1}{\delta})}$ which yield a sampling complexity :

$$T\geq\left(\frac{c}{\varepsilon}\right)^{2}\log(\frac{1}{\delta})$$

Given that the best known regret bounds have a dependence of $\frac{e^{\beta H}-1}{\beta}$, any sampling complexity obtained by this method will have a sampling complexity of $\left(\frac{e^{\beta H}-1}{\beta}\right)^{2}$.
Remark that if we have only a result on the expectation of the regret. Using Markov's inequality will yeild a sampling complexity in $\frac{1}{\delta^2}$ as mentioned in Ménard et al. (2021) which is undesirable.

3 The Exponential Curse: Lower Bounds for the Entropic Risk Measure

In the context of Best policy identification, establishing lower bounds consists of finding the smallest number of trajectories any algorithm must collect to achieve an (ε,δ)-PAC guarantee. Equivantely, if fewer than n trajectories are used, we are able to construct an instance on which the algorithm fails.

To establish such bounds, the main idea is to construct a very hard example that requires a very accurate knowledge of the environment to hope to estimate the optimal value function, one approach to prove (ε,δ)–PAC lower bounds for Best-Policy Identification, we construct a *hard pair* of MDPs $(\mathcal{M}^{+},\mathcal{M}^{-})$ such that: (i) they are *statistically close*, so that n trajectories do not easily reveal which environment generates the data; yet (ii) they have *different*

optimal policies and induce a return value gap of at least 2ε between these policies.These two requirements seems contradictive at first glance, but striking a balance between them is what will allow us to establish meaningful lower bounds. Any algorithm that cannot reliably tell $\mathcal{M}^+$ from $\mathcal{M}^-$ will, on at least one instance, select a policy whose value is worse than the optimal by more than ε with probability $> \delta$. This is what called in statistical learning LeCam's approach

3.1 LeCam's two points method and Information theory

LeCam's method is used widely in statistical learning theory to establish lower bounds on learning problems by reducing estimation to a binary hypothesis test between two carefully chosen and hard to distinguish data-generating models.
More formally, consider a measurable space $(\mathcal{X}, \mathcal{A})$ and for each parameter$\theta \in \Theta$, let P_θ be a probability measure on $(\mathcal{X}, \mathcal{A})$ and denote $P_\theta^{\otimes n}$ be the data law for n i.i.d samples. And let $\rho : \Theta \times \Theta \to \mathbb{R}^+$ be a metric.

LeCam's method establish a lower bound on the probability of error when reducing the estimation to a binary hypothesis tests between two fixed parameters θ_0,θ_1 such that $\rho(\theta_0, \theta_1) \geq 2\varepsilon$ so that success for both simultaneously is impossible :

Theorem 3.1. *Fix θ_0,$\theta_1 \in \Theta$ such that $\rho(\theta_0, \theta_1) \geq 2\varepsilon$, then :*

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

The method consists of picking two parameters θ_0 and θ_1 and make distinguish between them very hard by making the total variation distance between $\|P_{\theta_0}^{\otimes n} - P_{\theta_1}^{\otimes n}\|_{\mathrm{TV}}$ small and then solve the inequality given by the inequality above.

In practice making the total variation distance small is not very convenient so instead we make use of results from information theory.

3.2 Lower bound for entropic risk measure

For the entropic risk criterion, the exponential tilting $e^{\beta R}$ amplifies tail trajectories—upweighting rare, high-return paths when $\beta > 0$ (or high-cost paths in cost form) so if the optimal performance relies on a hard state–action–time pair $(s^\star, a^\star, h^\star)$ with tiny reachability, the learner must repeatedly realize that rare transition for the exponentially weighted estimates to stabilize, which inflates estimator variance and induces hardness that grows roughly like $e^{|\beta|H}$, making learning strictly harder than in the risk-neutral case.

To establish a lower bound, we use LeCam's approach. We will build two hard instances of an MDP that are very hard to distinguish by any algorithm in the sense that the KL divergence of the law of n episodes is small. We make the optimal policy differ by one action between the two MDPs then use the Bretagnolle–Huber inequality to express how hard it is to chose the correct optimal policy.

Using this approach we prove a lower bound on the best policy identification problem :

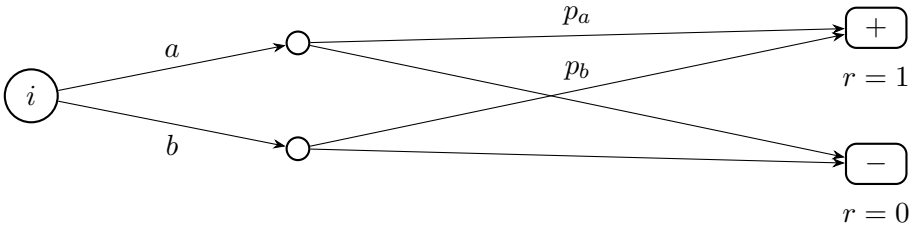

Figure 2: A simple two-armed bandit decision process. Node i represents the initial state. Two actions a and b lead to stochastic outcomes represented by the small chance nodes. Terminal outcomes $+$ and $-$ yield rewards $r = 1$ and $r = 0$ respectively, with probabilities p_a and p_b for reaching $+$.

Theorem 3.2. *Any algorithm that solves the (ε, δ)-PAC problem for ε satisfying $\beta\varepsilon \leq \ln(2)$ must have a sampling complexity that satisfy :*

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

sketch of the proof. Like we said we will construct two hard instances of an MDP that are hard to distinguish and using results from information theory, we will convert this hardness to a sampling complexity. Consider the following bandit problem: From state s_0 Chooses either action $u \in \{a, b\}$ and get the reward :

$$X_u = \begin{cases} R & \text{with probability } p_u \\ 0 & \text{with probability } 1 - p_u \end{cases}$$

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

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

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

$$V(p) - V(q) = \frac{1}{\beta} \log\left(\frac{1 + \lambda c p}{1 + \lambda c q}\right) = \frac{1}{\beta} \log(1 + \frac{\eta}{3})$$

Hence, we chose

$$\eta = 3(e^{\beta\varepsilon} - 1)$$

The hardness of distinguishability between those two MDPs is expressed in terms of the KL divergence. Let $\mathbb{P}^+_{1:n}$ and $\mathbb{P}^-_{1:n}$ denote the laws of the entire n-episode simulation under $\mathcal{M}^+$ and $\mathcal{M}^-$, respectively, we get :

$$\mathrm{KL}\left(\mathbb{P}^+_{1:n} \middle\| \mathbb{P}^-_{1:n}\right) \leq n\lambda \frac{\eta^2}{2c-1} = n\frac{\eta^2}{2c-1}$$

Now using 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 :

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

solving this we get

$$n \;\geq \frac{2c-1}{\eta^2}\log\frac{1}{4\delta} \qquad c = e^{\beta R} - 1\eta = 3(e^{\beta\varepsilon} - 1) \quad \text{and } (F)\text{holds}$$

Moreover if $\beta\varepsilon \leq \ln(2)$ then :

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

□

Let us try to extend the lower bound to be valid for any MDP. The idea to introduce S is to have S states and after the initial state the next state is drawn uniformly from $\{1, ..., |S|-3\}$. After drawing the state we are back to a bandit problem that is the same as the theorem above. After reaching a state at step 2 we stay there for all remaining $H-2$ steps and these introduce the H dependence and finally we introduce A by having the agent choose one of A action. We get the following lower bound :

Theorem 3.3 (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*

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

proof sketch. Inspired by Dann and Brunskill (2015). We introduce the following MDP: We start at state s_0. We pick a random number uniformly from $\{1, ..., n\}$ where $n = S - 3$ and we jump to bandit i. The learned take an action $a \in \{1, ..., A\}$ to jump either to $+$ to receive reward 1 or jump to $-$ and receive reward 0 and stays there for the rest of the episode. That means that if we take an optimal choice we receive a reward $H - 2$ at the end, otherwise we receive 0.

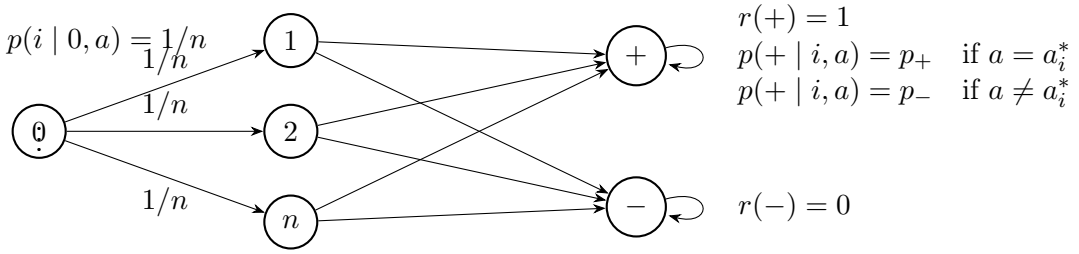

Figure 3: Representation of an MDP consisting of n parallel bandit problem. The initial state 0 transitions uniformly to one of the n bandits $i \in \{1, 2, \ldots, n\}$. Each bandit i can then lead to terminal states $+$ (reward 1) or $-$ (reward 0) depending on the chosen action. Probabilities p_+ and p_- indicate the success likelihood depending on whether the optimal action a_i^* is chosen. The terminal states are absorbing.

For each bandit i there is one optimal arm a_i^* that takes us to $+$ with probability p_+ and $A-1$ arms that takes us to $-$ with probability p_-. Which arm is optimal for i define a vector $I \in [A]^n$. We constructed a family of MDPs and each MDP is identifiable by the vector I i.e which arms are optimal for each bandit.

The bandit-level entropic value function is :

$$V(p) = \frac{1}{\beta} \log(1 + cp)$$

Where $c = e^{\beta(H-2)} - 1$. We choose p_+ and p_- in the same style in the proof above, so that the entropic gap is at most ε we get :

$$p_- = \frac{1}{2c} \qquad p_+ = p_-(1 + \eta) \qquad \eta = 3(e^{\frac{\beta \varepsilon_{\text{step}}}{H}} - 1)$$

Then we prove that a policy is ε-optimal, the policy need at least to chose an optimal arm in a fraction of the bandits. We try then to solve the different bandits with equal probability and the number of steps to do that provide a lower bound. □

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

4.1 Planning for the Bellman equation in a KL confidence region

Consider an finite episodic MDP $\Big(\mathcal{S}, \mathcal{A}, H, \{p_h\}_{h\in[H]}, \{r_h\}_{h\in H}\Big)$. Assume the rewards are deterministic, bounded in [0,1].
Let $n_h^t(s,a,s') = \sum_{i=1}^t \mathbf{1}\{(s_h^i, a_h^i, s_{h+1}^i) = (s,a,s')\}$ and $n_h^t(s,a) = \sum_{i=1}^t \mathbf{1}\{(s_h^i, a_h^i) = (s,a)\}$ and we define the empirical transition probability as :

$$\widehat{p}_h^t(s'|s,a) = \frac{n_h^t(s,a,s')}{n_h^t(s,a)} \quad \text{if} \quad n_h^t(s,a) \neq 0 \quad \text{and} \quad \widehat{p}_h^t(s'|s,a) = \frac{1}{S} \quad \text{otherwise}$$

Planning with a confidence region is a popular approach to do learning in reinforcement learning with finite time guarantees. Since planning is usually well understood, the method consist of doing planning in a learned MDP that has empirically constructed transition kernels that lies within a confidence region of the true transition probabilities. This yield bonus term of the form :

$$\overline{Q}_h^{t,\pi}(s,a) \triangleq \big(r_h + \max_{\bar{p}_h \in \mathcal{C}_h^t(s,a)} \bar{p}_h \overline{V}_{h+1}^{t,\pi}\big)(s,a), \qquad \underline{Q}_h^{t,\pi}(s,a) \triangleq \big(r_h + \min_{p_h \in \mathcal{C}_h^t(s,a)} p_h \underline{V}_{h+1}^{t,\pi}\big)(s,a),$$

$$\overline{V}_h^{t,\pi}(s) \triangleq \pi \overline{Q}_h^{t,\pi}(s), \qquad \underline{V}_h^{t,\pi}(s) \triangleq \pi \underline{Q}_h^{t,\pi}(s),$$

$$\overline{V}_{H+1}^{t,\pi}(s) \triangleq 0, \qquad \underline{V}_{H+1}^{t,\pi}(s) \triangleq 0,$$

$$\bar{p}_h^{t,\pi}(s,a) \in \arg\max_{\bar{p}\in\mathcal{C}_h^t(s,a)} \bar{p}\overline{V}_{h+1}^{t,\pi}, \qquad p_h^{t,\pi}(s,a) \in \arg\min_{p\in\mathcal{C}_h^t(s,a)} p\underline{V}_{h+1}^{t,\pi}.$$

Many confidence regions have used in the literature: ℓ^1-confidence regions using bounds on the TV distance between the true transition probability and its statistical estimation. Ménard et al. (2021) use D_{KL} divergence confidence region :

$$\mathcal{C}_h^t(s,a) \triangleq \left\{ q \in \Sigma_S : \mathrm{KL}\big(\widehat{p}_h^t(s,a), q\big) \leq \frac{\beta\big(n_h^t(s,a), \delta\big)}{n_h^t(s,a)} \right\}$$

The KL-divergence is very adapted to concentrate the estimated transition probability around the true one, since controlling at each state-action pair at each time step allows us to control the KL divergence of the whole trajectory distribution via the decomposition :

$$\mathrm{D}_{KL}\big(\widehat{P}^{t,\pi}, P^\pi\big) = \sum_{h=1}^H \sum_{s,a} \widehat{p}_h^{t,\pi}(s,a) \mathrm{KL}\big(\widehat{p}_h^{t,\pi}(\cdot \mid s,a) \| p_h(\cdot \mid s,a)\big)$$

4.2 Bonus terms

Motivated by the same approach as Ménard et al. (2021). We define bonuses term based on the Bernstein-inequality in Lemma 10. which converts the KL confidence regions to variance

aware upper bounds: An immediate application of lemma 10 yield an ideal bonus term that is not computable in practice since we do not know V^* :

$$B_h^t(s,a) \;\triangleq\; \sqrt{2\mathrm{Var}_{p_h}\big(e^{\beta V_{h+1}^\star}\big)(s,a)\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3e^{\beta H}\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$

And an associated shrinkage term to keep count of the local-lipschtizity of the log : and the associated *shrinkage factor* for the local Lipchitz control of $\log(\cdot)$ is

$$s_h^t(s,a) \;\triangleq\; \frac{1}{\max\big\{1,\; \widehat{p}_h^t V_{h+1}^\star(s,a) - B_h^t(s,a)\big\}}$$

Using lemma 11 to transport from the variance under V^* to the computable variance $\widehat{V}$ and derive computable upper bounds:

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

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

The bonus we have is decomposed to two terms, the first one is a variance aware bonus that decays in $\frac{1}{\sqrt{n}}$ while the second term is a range term that decays in $\frac{1}{n}$. Using these bonuses we can define upper and lower Q- and value functions:

$$\begin{aligned}
\widetilde{Q}_h^t(s,a) \triangleq \min\Big\{H, r_h(s,a) &+ \frac{1}{\beta}\log\big(1+\widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)\big) + \frac{1}{\beta H}\widehat{p}_h^t\big(e^{\beta\widetilde{V}_{h+1}^t} - e^{\beta\underline{V}_{h+1}^t}\big)(s,a) \\
&+\; \rho_\beta^{\widehat{p}_h^t}\big(\widetilde{V}_{h+1}^t\big)(s,a)\Big\}, \\
\underline{Q}_h^t(s,a) \triangleq \max\Big\{0, r_h(s,a) &- \frac{1}{\beta}\log\big(1+\widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)\big) - \frac{1}{\beta H}\widehat{p}_h^t\big(e^{\beta\widetilde{V}_{h+1}^t} - e^{\beta\underline{V}_{h+1}^t}\big)(s,a) \\
&+\; \rho_\beta^{\widehat{p}_h^t}\big(\underline{V}_{h+1}^t\big)(s,a)\Big\}, \\
\widetilde{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}}\widetilde{Q}_h^t(s,a), &\qquad \underline{V}_h^t(s) \triangleq \max_{a\in\mathcal{A}}\underline{Q}_h^t(s,a), \qquad \widetilde{V}_{H+1}^t = \underline{V}_{H+1}^t \equiv 0.
\end{aligned}$$

The upper/lower bounds we established verify the optimism lemma :

Lemma 4.1 (Optimism and pessimism, entropic case)**.** *On $\mathcal{G}$, for all $t \geq 0$, $h \in [H]$, and (s,a),*

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

Next, We also build upper bounds on the gap $V_1^*(s) - V_1^{\pi^{t+1}}(s_1)$ in order to establish a stopping rule. We define G recursively as :

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

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

Input: $\beta > 0$, $\delta \in (0,1)$, $\varepsilon > 0$.
Initialize counts $n_h^0(\cdot) = 0$ and $\widehat{p}_h^0(\cdot|s,a) = 1/S$.
for $t = 0, 1, 2, \dots$ **do**
  Compute $\widetilde{B}_h^t$ and $\widetilde{s}_h^t$; then $(\widetilde{Q}^t, \widetilde{V}^t)$ and $(\underline{Q}^t, \underline{V}^t)$.
  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

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

  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

This allow us to derive a UCB-type algorithm for the BPI for entropic risk measure : We prove then that for the greedy algorithm π^{t+1} we have :

Lemma 4.2. *The greedy policy satisfies the standard gap domination:*

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

This algorithm allow us to obtain a sampling complexity for the entropic best policy identification :

Theorem 4.1. *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:*

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

Proof sketch. Using lemma 10 and the inequality $\mathrm{Var}_{p_h} \le (\pi_{h+1}^{t+1} G_{h+1}^t(s,a)) \le H p_h \pi_{h+1}^{t+1} G_{h+1}^t(s,a)$ and the inequality $\sqrt{xy} \le x + y$ we have :

$$\Big(\widehat{p}_h^t - p_h\Big)\pi_{h+1}^{t+1} G_{h+1}^t(s,a) \le \frac{1}{H} p_h \pi_{h+1}^{t+1} G_{h+1}^t(s,a) + 3H^2 \frac{\beta(n_h^t(s,a), \delta)}{n_h^t(s,a)}$$

Now we use the inequality $\log(1+x) \le x$ (note that keeping the log term is tighter when we are still early in the algorithm but as $\widehat{p}$ gets close to p we can get rid of the log to simplify

the analysis), we get then :

$$G_h^t(s,a) \leq \frac{6}{\beta}\sqrt{\mathrm{Var}_{p_h}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a)\ \Big(\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\Big)} + \frac{36}{\beta}e^{\beta H}H^2\Big(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\Big) \\ + \ \Big(1+\frac{13}{H}\Big)p_h\pi_{h+1}^{t+1}G_{h+1}^t(s,a)$$

Unfolding the recursion and using $\big(1+\frac{13}{H}\big)^H \leq e^{13}$ we get the single-episode bound :

$$\pi_1^{t+1}G_1^t(s_1) \leq \frac{6}{\beta}e^{13}\sum_{h=1}^H\ \sum_{s,a} p_h^{t+1}(s,a)\sqrt{\mathrm{Var}_{p_h}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a)\ \Big(\frac{\beta^\star(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\Big)} \\ + \ \frac{36}{\beta}e^{13}H^2\sum_{h=1}^H\ \sum_{s,a} p_h^{t+1}(s,a)\Big(\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \wedge 1\Big)$$

Finally, we switch counts to pseudo-counts by Lemma 8 (on $\mathcal{E}_{\mathrm{cnt}}$), which multiplies the fractions by at most 4 and replaces n_h^t by $\bar{n}_h^t$ (the expected visitation counts), and use the fact that for any $t \leq T$ we have $\pi_1^{t+1}G_1^t(s_1) \geq \varepsilon$ so summing the inequality over $t = 1, ..., T$ and using Cauchy-Schwartz gives:

$$(T+1)\varepsilon \leq \frac{24}{\beta}e^{13}\sqrt{T+1}\ \sqrt{\sum_{t=0}^T\sum_{h,s,a} p_h^{t+1}(s,a)\mathrm{Var}_{p_h}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a)}\ \sqrt{\sum_{t=0}^T\sum_{h,s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t,\delta)}{\bar{n}_h^t \vee 1}} \\ + \frac{144}{\beta}e^{13}H^2\sum_{t=0}^T\sum_{h,s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t,\delta)}{\bar{n}_h^t \vee 1}$$

We control the first term using the elementary inequality $\mathrm{Var}_{p_h}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a) \leq \frac{(e^{\beta H}-1)^2}{4}$. Remark here that contrary to Ménard et al. (2021), we can't use a total law of variance since the entropic risk measure is not sub-additive.
Using the psuedo-counts lemma from Ménard et al. (2021) we have :

$$\sum_{t=0}^T\sum_{h,s,a} p_h^{t+1}(s,a)\frac{\beta^\star(\bar{n}_h^t,\delta)}{\bar{n}_h^t \vee 1} \leq \beta^\star(T,\delta)\sum_{h,s,a}\sum_{t=0}^T \frac{\bar{n}_h^{t+1}(s,a)-\bar{n}_h^t(s,a)}{\bar{n}_h^t(s,a)\vee 1} \ \leq \ 4HSA\beta^\star(T,\delta)\log(T+2),$$

$$\sum_{t=0}^T\sum_{h,s,a} p_h^{t+1}(s,a)\frac{\beta(\bar{n}_h^t,\delta)}{\bar{n}_h^t \vee 1} \leq 4HSA\beta(T,\delta)\log(T+2),$$

Hence we get the bound :

$$(T+1)\varepsilon \leq 24\frac{e^{\beta H}-1}{\beta}\sqrt{T+1}\ \sqrt{H^2SA\beta^\star(T,\delta)\log(T+2)} + 144\frac{e^{\beta H}-1}{\beta}H^3SA\beta(T,\delta)\log(T+2)$$

This is true for any $T < \tau$, hence using the expressions of β and β^* we get the bound :

$$\tau\varepsilon \leq 24\frac{e^{\beta H}-1}{\beta}\sqrt{\tau H^2 SA\Big(\log(3SAH/\delta)\log(8e\tau)+\log(8e\tau)^2\Big)} + 144\frac{e^{\beta H}-1}{\beta}H^3SA\Big(\log(3SAH/\delta)\log(8e\tau)+S\log(8e\tau)^2\Big)$$

Use lemma 13 to solve this inequality yield the upper bound with probability $1-\delta$:

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

□

5 A promising approach

Bernstein's inequality naturally involves the variance of $e^{\beta V}$ when applying it to the entropic risk measure; this introduces a factor of $e^{2\beta H}$ in the sample complexity. This suggests that, to obtain optimal rates, we need a more appropriate notion of sample complexity. One idea is to define the bonus terms using the tilted variance instead of $\mathrm{Var}(e^{\beta V})$. More precisely, the tilted variance is defined as :

$$\frac{dP_\beta}{dP}(v) = \frac{e^{\beta v}}{\mathbb{E}[e^{\beta V}]}$$

The tilted mean and variance are

$$\mu_\beta = \mathbb{E}_\beta[V], \qquad \sigma_\beta^2 = \mathrm{Var}_\beta(V) = \mathbb{E}_\beta\big[(V-\mu_\beta)^2\big]$$

And

$$\rho'(\beta) = \mathbb{E}_\beta[V] = \mu_\beta, \qquad \rho''(\beta) = \mathrm{Var}_\beta(V) = \sigma_\beta^2$$

Thus, σ_β^2 is precisely the local curvature of the entropic risk. Intuitively, the tilted variance guides exploration toward high-reward regions for risk-seeking agents and low rewards for risk-averse agents while avoiding the blow-up induced by the exponential transform $e^{\beta V}$.
For now we have the result :

Lemma 5.1. *Let p, q be two probability measures on $\mathcal{S}$ and let $f : \mathcal{S} \to [0, b]$ then for any $\lambda > 0$:*

$$sgn(\beta)(\rho_{\beta,p}(f) - \rho_{\beta,q}(g)) \leq \frac{\lambda}{|\beta|(1+\lambda)}D_{1+\lambda}(p||q) + \frac{\lambda}{|\beta|(1+\lambda)}\frac{Var_{q_\beta}(f)}{R^2}\phi(\frac{|\beta|H}{\lambda})$$

Where $\phi(x) = e^x - x - 1$ is the Bennett function and D_α is the Renyi divergence defined as :

$$D_\alpha(P\|Q) = \frac{1}{\alpha-1}\log\sum_i p_i^\alpha q_i^{1-\alpha}.$$

It is a generalization for the KL divergence since $D_\alpha \to D_{KL}$ as $\alpha \to 1$.
To use the same approach, we need to build confidence balls for the Renyi divergence for $\alpha > 1$, this is unfortunately is not possible as states the lemma :

Theorem 5.1. *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 $b(n,\delta) = O(\log(1/\delta))$

One possible fix, that I still need to work out its details is to have a burn-in phase so that we have a well-behaved Renyi divergence :

Theorem 5.2 (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 $p_{\min} \doteq \min_i p_i \geq b > 0$. 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

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

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

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

6 Appendix

6.1 Proofs from section 2

6.1.1 Results from Information theory

In information theory, the Kullback-Leilber divergence is a pseudo-statistical distance that measure how much a probability distribution Q is different from another probability distribution P that is defined from Shannon's entropy, mathematically we write :

Definition 2. *Let $(\Omega, \mathcal{F})$ be a measurable space and let P and Q be two probability measures on $(\Omega, \mathcal{F})$, we define the KL divergence as :*

$$D_{KL}(P,Q) = \mathbb{E}_P\left[\log \frac{p(X)}{q(X)}\right] = \begin{cases} \int \log\left(\frac{dP}{dQ}(\omega)\right) dP(\omega) & \text{if } P \ll Q \\ \infty, & \text{otherwise.} \end{cases}$$

Intuitively, $D(P\|Q)$ is the expected extra codelength (in bits) when data come from P but we encode with a code matched to Q. Or equivalently, it's the average log-likelihood ratio—how distinguishable P is from Q per symbol.
One frequently used example is when P, Q are both Bernoulli random variables, we denote it as $d(p, q)$:

$$D_{KL}(\mathcal{B}(p), \mathcal{B}(q)) = d(p,q) = p\log(\frac{p}{q}) + (1-p)\log(\frac{1-p}{1-q})$$

KL divergence is a pseudo-distance in the sense that it is positive definite but it is not symmetric nor verify the triangle inequality but it's still a very usefull tool to compare two probability distributions that is easily controllable. In particular, we have many results from information theory that shows the worth of working with KL divergence.

Theorem 6.1 (Pinsker's inequality)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let P, Q be probability measures on it. Then*

$$\|P - Q\|_{TV} \leq \sqrt{\frac{1}{2} D_{KL}(P\|Q)},$$

We also have the Bretagnolle–Huber inequality :

Theorem 6.2 (Bretagnolle–Huber inequality)**.** *Let $(\Omega, \mathcal{F})$ be a measurable space and let P, Q be probability measures on it, and let $A \in \mathcal{F}$, Then*

$$P(A) + Q(A^c) \geq \frac{1}{2}\exp(-D_{KL}(P\|Q))$$

In particular by taking the event $A = \{\frac{dP}{d\mu} \geq \frac{dQ}{d\mu}\}$ (where both P and Q are absolutely continuous with respect to μ, for example $\mu = \frac{P+Q}{2}$),

$$\|P - Q\|_{TV} \leq 1 - \frac{1}{2}\exp(-D_{KL}(P\|Q))$$

Both of these inequalities shows that we can use the KL divergence to build indistinguishable models instead of the total variance distance, this is very nice since in Reinforcement learning we have the decomposition :

Lemma 6.1 (KL decomposition for MDPs)**.** *Let $M = (\mathcal{S}, \mathcal{A}, P, R, \mu_0)$ and $M' = (\mathcal{S}, \mathcal{A}, P', R', \mu_0)$ be two MDPs. Fix a horizon H and a policy π. Let $\mathbb{P}_M^\pi$ and $\mathbb{P}_{M'}^\pi$ be the trajectory measures on $(\mathcal{S} \times \mathcal{A} \times \mathbb{R})^n$ induced by simulating π with M and M', respectively.Then*

$$D_{KL}(\mathbb{P}_M^\pi \| \mathbb{P}_{M'}^\pi) = \mathbb{E}_{M,\pi}\left[\sum_{t=0}^{n-1} D_{KL}(P(\cdot \mid S_t, A_t) \| P'(\cdot \mid S_t, A_t)) + D_{KL}(R(\cdot \mid S_t, A_t) \| R'(\cdot \mid S_t, A_t))\right]$$

6.1.2 Proof of theorem 3

Proof. Consider the following MDP : Fix $R > 0$ and the horizon $H = 1$ (bandit problem)

- **States** Start, Left, Right, Terminal
- **Step** $h = 0$:
  - At right, terminate with reward 0
  - Chooses either action $u \in \{a, b\}$ and get the reward :

$$X_u = \begin{cases} R & \text{with probability } p_u \\ 0 & \text{with probability } 1 - p_u \end{cases}$$

In this environment a policy is identified by picking up action a or b, hence we can express the entropic value function as a function of p, if we denote $c = e^{\beta R} - 1$ then :

$$G(p) = \frac{1}{\beta} \log(1 + cp)$$

Where p is either p_a or p_b. The MDPs is identified by given p_a and p_b.
We now define two MDPs $\mathcal{M}^+$ and $\mathcal{M}^-$ within the family of MDPs given above that differs 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

$$\eta = 3(e^{\beta\varepsilon} - 1)$$

To have both $p, q \in [0, 1]$ we need :

$$2c \geq 1 + \eta = 3e^{\beta\varepsilon} - 2$$

Which is verified for ε small enough.
Now we have $G(p) - G(q) = \varepsilon$. This means that on $\mathcal{M}^+$, action a is optimal and beats b by a gap ε while on $\mathcal{M}^-$ it's the opposite, b beats a with a gap ε.
Let $\mathbb{P}^+_{1:n}$ and $\mathbb{P}^-_{1:n}$ denote the laws of the entire n-episode simulation under $\mathcal{M}^+$ and $\mathcal{M}^-$, respectively, for a learner. By the chain rule for KL and conditioning on whether **Left** is reached,

$$\mathrm{KL}\left(\mathbb{P}^+_{1:n} \| \mathbb{P}^-_{1:n}\right) = \sum_{t=1}^{n} \mathbb{E}\big[\mathrm{KL}(\text{episode } t \mid \text{history})\big] \leq \ nJ$$

where

$$J = \max\{d(p, q), d(q, p)\}$$

Since $\mathrm{KL}(Q\|P) \leq \chi^2(Q\|P)$ and $\chi^2(\mathrm{Bern}(q)\|\mathrm{Bern}(p)) = \frac{(q-p)^2}{p(1-p)}$ we have

$$J_{\max} \leq \frac{(p_+ - p_-)^2}{p_-(1-p_-)} = \frac{(p_-\eta)^2}{p_-(1-p_-)} = \frac{p_-}{1-p_-}\eta^2$$

$$J \leq \frac{(p-q)^2}{q(1-q)} = \frac{(q\eta)^2}{q(1-q)} = \frac{q}{1-q}\eta^2 = \frac{\eta^2}{2c-1}$$

Hence :

$$\mathrm{KL}\left(\mathbb{P}^+_{1:n} \middle\| \mathbb{P}^-_{1:n}\right) \leq n\frac{\eta^2}{2c-1}$$

Now 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 :

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

solving this we get

$$n \ \geq \ \frac{2c-1}{\eta^2}\log\frac{1}{4\delta} \qquad c = e^{\beta R} - 1 \ \ \eta = 3(e^{\beta\varepsilon} - 1) \ \text{ and } (F) \text{ holds}$$

Moreover if $\beta\varepsilon \leq \ln(2)$ then :

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

□

Proof. Inspired by Dann and Brunskill (2015). We introduce the following MDP:
We start at state s_0. We pick a random number uniformly from $\{1, ..., n\}$ where $n = S - 3$ and we jump to bandit i. The learned take an action $a \in \{1, ..., A\}$ to jump either to $+$ to receive reward 1 or jump to $-$ and receive reward 0 and stays there for the rest of the episode. That means that if we take an optimal choice we receive a reward $H - 2$ at the end, otherwise we receive 0.

For each bandit i there is one optimal arm a_i^* that takes us to $+$ with probability p_+ and $A-1$ arms that takes us to $-$ with probability p_-. Which arm is optimal for i define a vector $I \in [A]^n$. We constructed a family of MDPs and each MDP is identifiable by the vector I i.e which arms are optimal for each bandit.

The bandit-level entropic value function is :

$$V(p) = \frac{1}{\beta} \log(1 + cp)$$

Where $c = e^{\beta(H-2)} - 1$. We choose p_+ and p_- in the same style in the proof above, so that the entropic gap is at most ε we get :

$$p_- = \frac{1}{2c} \qquad p_+ = p_-(1 + \eta) \qquad \eta = 3(e^{\frac{\beta \varepsilon_{\text{step}}}{H}} - 1)$$

Then we prove that a policy is ε-optimal, the policy need at least to chose an optimal arm in a fraction of the bandits. We try then to solve the different bandits with equal probability and the number of steps to do that provide a lower bound.

Now, for a deterministic policy π, let M^π be the number of bandits on which π picks a suboptimal arm and let $\rho = \frac{M}{n}$. The optimal policy picks the optimal arm for each bandit, hence :

$$V^* = \frac{1}{\beta} \log(1 + cp_+)$$

And:

$$V^\pi = \frac{1}{\beta} \log \left(\frac{(n - M)(1 + cp_+) + M(1 + cp_-)}{n} \right) = \frac{1}{\beta} \log \Big((1 + cp_+)(1 - \rho + \rho e^{-\beta \varepsilon_{\text{step}}}) \Big)$$

Since $e^{-\beta \varepsilon_{\text{step}}} = \frac{1+cp_+}{1+cp_-}$. Hence, the optimality gap verify :

$$V^* - V^\pi = \frac{-1}{\beta} \log \Big(1 - (1 - r)\rho \Big)$$

If π is ε-optimal then :

$$\rho \leq \frac{1 - e^{-\beta \varepsilon}}{1 - e^{-\beta \varepsilon_{\text{step}}}}$$

This means the policy must at least solve a fraction $\phi = 1 - \frac{1-e^{-\beta\varepsilon}}{1-e^{-\beta\varepsilon_{\text{step}}}}$ of the bandits in the MDP to be ε-optimal. Enforce this with a union bound by requiring the per-bandit error to satisfy $\delta_i \leq \delta/(\phi n)$ for those ϕn bandits.

Fix such a bandit i and the true instance where its unique optimal arm is $a_i^\star$ with success probability p_+. For each suboptimal arm $b \neq a_i^\star$, define an alternative MDP $\nu^{(b)}$ that changes only arm b to be the optimal arm. Applying lemma 1 from (Kaufmann et al., 2016) to the event E_b 'the algorithm recommends arm b for bandit i' ,since the bandit is solved correctly, we have $\Pr(E_b)$, under instance v, recommending b is a mistake hence $\Pr_v(E_b) \leq \delta_i)$ and since b is optimal under $v^{(b)}$, recommending b is correct hence gives $\Pr_{v^{(b)}}(E_b) \geq 1 - \delta_i)$:

$$\mathbb{E}[N_{i,b}] \geq \frac{D_{KL}\big(1 - \frac{\delta}{\phi n}, \frac{\delta}{\phi n}\big)}{D_{KL}\big(\text{Ber}(p_-)\|\text{Ber}(p_+)\big)}$$

Summing over the $A-1$ suboptimal arms b,

$$\mathbb{E}[N_i] = \sum_{a=1}^{A} \mathbb{E}[N_{i,a}] \geq (A-1)\frac{D_{KL}\big(1 - \frac{\delta}{\phi n}, \frac{\delta}{\phi n}\big)}{D_{KL}\big(\text{Ber}(p_-)\|\text{Ber}(p_+)\big)}.$$

Finally, each episode contributes exactly one pull in exactly one bandit, so $T = \sum_{i=1}^{n} \mathbb{E}[N_i]$. Summing over the ϕn bandits that must be solved and using $D_{KL}(1-x, x) \geq \log(1/(2x))$(assuming $\delta \leq \frac{n}{2}$ yields :

$$T \geq \phi(S-3)(A-1)\frac{\log\big(\frac{\phi(S-3)}{2\delta}\big)}{D_{KL}\big(\text{Ber}(p_-)\|\text{Ber}(p_+)\big)}$$

And since $D_{KL}\big(\text{Ber}(p_-)\|\text{Ber}(p_+)\big) \leq \frac{(p_- - p_+)^2}{p_-(1-p_-)} \leq \frac{\eta^2}{2c-1}$ We get :

$$T \geq 4\frac{e^{2(\beta(H-2)} - 1) - 1}{(e^{\beta\varepsilon_{\text{step}}} - 1)^2}\phi(S-3)(A-1)\log\big(\tfrac{\phi(S-3)}{2\delta}\big)$$

We can pick for example $\varepsilon_{\text{step}} = 2H\varepsilon$ so that for ε small enough, $\phi \geq 2$ for small ε :

$$T = \mathcal{O}\Big((S-3)(A-1)\frac{2(e^{\beta(H-2)} - 1) - 1}{\beta^2\varepsilon^2}\log(S/\delta)\Big)$$

□

6.2 Proofs of section 3

In the case where we want to maximize the entropic risk measure $\rho_\beta(X) = \log(\mathbb{E}[e^{\beta X}])$ instead of the expectation we have :

$$Q_h^\pi(s,a) \triangleq \rho_\beta\Big(r_h(s,a) + V_{h+1}^\pi(s')\Big) = \frac{1}{\beta}\ln\left(\sum_{s'\in S} p_h(s',a)\exp\big[\beta\big(r_h(s,a)+V_{h+1}^\pi(s')\big)\big]\right)$$
$$= \frac{1}{\beta}\ln(p_h[e^{r_h(s,a)+V_{h+1}^\pi}](s,a)) = \rho_\beta^{(s,a)}(r_h(s,a)+V_{h+1}^\pi) \quad \text{where we take the expectation with respec}$$

$$V_h^\pi(s) \triangleq Q_h^\pi\big(s,\pi_h(s)\big), \qquad V_{H+1}^\pi(s) \triangleq 0$$

$$V_h^\star(s) \triangleq \max_{a\in A} Q_h^\star(s,a), \qquad V_{H+1}^\star(s) \triangleq 0$$

Let us try to build confidence bounds for this framework following the same trick, we define similar bounds for Q :

So that we can then prove following the same proof that :

Lemma 6.2. *On the event $\mathcal{G}$ we have :*

$$\underset{\sim}{Q}_h^t(s,a) \le Q_h^*(s,a) \le \widetilde{Q}_h^t(s,a)$$

This is the direct generalization of their bounds to the exponential case using a lose lipschtizness of the log, but we get a variance of an exponential transformation of $\widehat{V}_{h+1}^t$ which most certainly blows in an exponential of the horizon and β, what if we use the local lipschtizity of the log since the estimated function lies within a small interval of the true one with high probability ? Good track. Also How do we choose β^*.
Let us make use of the log :
Denote

$$\rho_\beta^p(X) = \frac{1}{\beta}\log(pe^\beta X)$$

Where p is the expectation. Denote the bonus term :

$$B_h^t(s,a) = \sqrt{2\mathrm{Var}_{p_h}(e^{\beta V_{h+1}^*})(s,a)\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3e^{\beta H}\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)}$$

And consider a shrinking factor :

$$s_h^t(s,a) = \frac{1}{\max\{1, \widehat{p}_h^t(s,a)V_{h+1}^* - B_h^t(s,a)\}}$$

$$\mathcal{E}^* = \{\forall t\in\mathbb{N}, \forall h\in[H], \forall (s,a)\in\mathcal{S}\times\mathcal{A} : |\rho_\beta^{\widehat{p}_h^t}(V_{h+1}^*(s,a)) - \rho_\beta^{\widehat{p}_h}(V_{h+1}^*(s,a))| \le \min(H, \log(1+s_h^t B_h^t(s,a))\}$$

Lemma 6.3. *With* $\beta^*(n,\delta) = \log(3HSA/\delta) + \log(8e(n-1))$ *we have :*

$$Pr(\mathcal{E}^*) \geq 1 - \delta/3$$

Proof. We will apply theorem 5 in the article to the exponential transform of V^*_{h+1} we get that with probability $1-\delta/3$:

$$|(\widehat{p}^t_h - p_h)e^{\beta V^*_{h+1}}| \leq \sqrt{2\mathrm{Var}_{p_h}(e^{\beta V^*_{h+1}})(s,a)\frac{\beta^*(n^t_h(s,a),\delta)}{n^t_h(s,a)}} + 3e^{\beta H}\frac{\beta^*(n^t_h(s,a),\delta)}{n^t_h(s,a)}$$

Now we can write :

$$|\log(p_h e^{\beta V^*_{h+1}}) - \log(\widehat{p}^t_h e^{\beta V^*_{h+1}})| = \log(1 + \frac{|(\widehat{p}^t_h - p_h)e^{\beta V^*_{h+1}}|}{\min(p_h e^{\beta V^*_{h+1}}, \widehat{p}^t_h e^{\beta V^*_{h+1}})})$$

Using the first inequality, we find that with probability $1 - /3$ that :

$$|\rho^{\widehat{p}^t_h}_\beta(V^*_{h+1}(s,a)) - \rho^{\widehat{p}_h}_\beta(V^*_{h+1}(s,a))| \leq \frac{1}{\beta}\log(1 + s^t_h(s,a)B^t_h(s,a))$$

And we conclude □

We aim then to prove a similar result to Lemma 5, First let us bound the bonus term by a calculable term, then we bound the variance using lemma 10 and 11 :

Lemma 6.4. *We have the bound*

$$\begin{aligned} Var_{p_h}(e^{\beta V^*_{h+1}}) &\leq 2\,Var_{\widehat{p}^t_h}(e^{\beta V^*_{h+1}})(s,a) + 4e^{2\beta H}\frac{\beta(n^t_h(s,a),\delta)}{n^t_h(s,a)} \\ &\leq 4\,Var_{\widehat{p}^t_h}(e^{\beta \widetilde{V}^t_{h+1}})(s,a) + 4e^{\beta H}\widehat{p}^h_t(e^{\beta \widetilde{V}^t_{h+1}} - e^{\beta V^*_{h+1}})(s,a) + 4e^{2\beta H}\frac{\beta(n^t_h,\delta)}{n^t_h(s,a)} \end{aligned}$$

And it follows that using the inequality $\sqrt{x+y} \leq \sqrt{x} + \sqrt{y}$:

$$
\begin{aligned}
B_h^t(s,a) &= \sqrt{2\mathrm{Var}(e^{\beta V_{h+1}^*})(s,a)\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3e^{\beta H}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\leq \sqrt{2\Big[4\mathrm{Var}_{\widehat{p}_h^t}(e^{\beta \widehat{V}_{h+1}^t})(s,a) + 4e^{\beta H}\widehat{p}_t^h(e^{\beta \widetilde{V}_{h+1}^t} - e^{\beta V_{h+1}^*})(s,a) + 4e^{2\beta H}\frac{\beta(n_h^t,\delta)}{n_h^t(s,a)}\Big]\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\
&+ 3e^{\beta H}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\leq 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(e^{\beta \widehat{V}_{h+1}^t})(s,a)\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + \sqrt{8e^{\beta H}\widehat{p}_t^h(e^{\beta \widetilde{V}_{h+1}^t} - e^{\beta V_{h+1}^*})\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} \\
&\qquad + \sqrt{8e^{2\beta H}\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)}} + 3e^{\beta H}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\leq 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(e^{\beta \widehat{V}_{h+1}^t})(s,a)}\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)} + \frac{1}{H}\widehat{p}_t^h(e^{\beta \widetilde{V}_{h+1}^t} - e^{\beta V_{h+1}^*}) + 8He^{\beta H}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\qquad + 2\sqrt{2}e^{\beta H}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} + 3e^{\beta H}\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&\leq 2\sqrt{2}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}(e^{\beta \widehat{V}_{h+1}^t})(s,a)}\frac{\beta^*(n_h^t(s,a),\delta)}{n_h^t(s,a)} + e^{\beta H}(8H + 2\sqrt{2} + 3)\frac{\beta(n_h^t(s,a),\delta)}{n_h^t(s,a)} \\
&+ \frac{1}{H}\widehat{p}_t^h(e^{\beta \widetilde{V}_{h+1}^t} - e^{\beta V_{h+1}^*})
\end{aligned}
$$

Let

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

Similarly, for the shrinking factor, we have :

$$
\widehat{p}_h^t(s,a)V_{h+1}^* - B_h^t(s,a) \geq \widehat{p}_h^t(s,a)V_{h+1}^* - \widetilde{B}_h^t(s,a) \geq \widehat{p}_h^t(s,a)\underline{V}_{h+1}^* - \widetilde{B}_h^t(s,a)
$$

Hence :

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

Hence :

$$
\log(1 + s_h^t(s,a)B_h^t(s,a)) \leq \log(1 + \widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)) + \log(1 + \frac{1}{H}\widehat{p}_t^h(e^{\beta \widetilde{V}_{h+1}^t} - e^{\beta V_{h+1}^*}))
$$

Where we used the inequality $\log(1+a+b) \leq \log(1+a) + \log(1+b)$ when $a, b \geq 0$ and that

$s_h^t(s,a) \leq 1$. We then define :

$$\widetilde{Q}_h^t(s,a) = \min(H, r(s,a) + \frac{1}{\beta}\log(1+\widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)) + \frac{1}{\beta H}\widehat{p}_t^h(e^{\beta\widetilde{V}_{h+1}^t} - e^{\beta\underset{\sim}{V}_{h+1}^t(s)}) + \rho_\beta^{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t(s))$$

$$\widetilde{V}_h^t(s) = \max_a \widetilde{Q}_h^t(s,a)$$

$$\underset{\sim}{Q}_h^t(s,a) = \max(0, r(s,a) - \frac{1}{\beta}\log(1+\widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)) - \frac{1}{\beta H}\widehat{p}_t^h(e^{\beta\widetilde{V}_{h+1}^t} - e^{\beta\underset{\sim}{V}_{h+1}^t(s))}) + \rho_\beta^{\widehat{p}_h^t}(\underset{\sim}{V}_h^t(s))$$

$$\underset{\sim}{V}_h^t(s) = \max_a \underset{\sim}{Q}_h^t(s,a)$$

We can then prove by induction that:

Lemma 6.5. *If $\beta^*(n,\delta) \leq \beta(n,\delta)$, then on the event $\mathcal{G}$ for any t and any $h \in [H]$ and all (s,a) :*

$$\underset{\sim}{Q}_h^t(s,a) \leq Q_h^*(s,a) \leq \widetilde{Q}_h^t(s,a)$$

And

$$\underset{\sim}{V}_h^t(s) \leq V_h^*(s) \leq \widetilde{V}_h^t(s)$$

$$\mathring{Q}_h^t(s,a) \triangleq \min(r_h(s,a) + \rho_\beta^{p_h}(\mathring{V}_h^t(s), \max(0, r(s,a) - \frac{1}{\beta}\log\Big(1+\tilde{s}_h^t(s,a)\tilde{B}_h^t(s,a)\Big)$$
$$- \frac{1}{\beta H}\widehat{p}_t^h(e^{\beta\widetilde{V}_{h+1}^t(s)} - e^{\beta\underset{\sim}{V}_{h+1}^t(s)}) + \rho_\beta^{\widehat{p}_h^t}(\mathring{V}_h^t(s))))$$

Proof. Let us try to upper bound $\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a)$:
First case, assume $\mathring{Q}_t^h(s,a) = r_h(s,a) + \rho_\beta^{p_h}(\mathring{V}_h^t(s)$:

$$\widetilde{Q}_h^t(s,a) - \mathring{Q}_h^t(s,a) \leq \frac{1}{\beta}\log(1+\widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)) + \frac{1}{\beta H}\widehat{p}_t^h(e^{\beta\widetilde{V}_{h+1}^t(s)} - e^{\beta\underset{\sim}{V}_{h+1}^t(s)})$$
$$+ (\rho_\beta^{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t(s)) - \rho_\beta^{p_h}(\mathring{V}_h^t(s)))$$

And we can decompose the last term as :

$$\rho_\beta^{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t(s)) - \rho_\beta^{p_h}(\mathring{V}_h^t(s)) = (\rho_\beta^{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t(s)) - \rho_\beta^{\widehat{p}_h^t}(\mathring{V}_{h+1}^t(s)))+$$
$$(\rho_\beta^{\widehat{p}_h^t}(V_{h+1}^*(s) - \rho_\beta^{p_h}(V_{h+1}^*(s))) + (\rho_\beta^{p_h} - \rho_\beta^{\widehat{p}_h^t})(V_{h+1}^* - \mathring{V}_{h+1}^t)$$

We need to bound the last term in this decomposition in function of $\widehat{p}_t^h(e^{\beta V_{h+1}^*(s)} - e^{\beta\mathring{V}_{h+1}^t(s)})$
Using Bernstein inequality we get :

$$\big(p_h - \widehat{p}_h^t\big)\big(e^{\beta V_{h+1}^*} - e^{\beta\mathring{V}_{h+1}^t}\big)(s,a) \leq \frac{1}{H}\widehat{p}_h^t\big(e^{\beta V_{h+1}^*} - e^{\beta\mathring{V}_{h+1}^t}\big)(s,a) + 8H^2e^{\beta H}\frac{\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}$$

Now notice that if we denote $\Delta_h^t(s,a) = V_{h+1}^* - \mathring{V}_{h+1}^t$ and $f = e^{\beta\Delta_h^t(s,a)}$ then :

$$\rho_\beta^{\widehat{p}_h^t}(V_{h+1}^* - \mathring{V}_{h+1}^t(s)) - \rho_\beta^{p_h}(V_{h+1}^* - \mathring{V}_{h+1}^t(s)) = \frac{1}{\beta}(\log(p_h f) - \log(\widehat{p}_h^t)) = \frac{1}{\beta}\log(1 + \frac{p_h f - \widehat{p}_h^t f}{\min\{p_h f, \widehat{p}_h^t f\}})$$
$$\leq \frac{1}{\beta}\log(1 + (p_h f - \widehat{p}_h^t f))$$

Hence using the first inequality and that $\log(1+a+b) \leq \log(1+a) + b$:

$$\begin{aligned}\rho_\beta^{\widehat{p}_h^t}(V_{h+1}^* - \mathring{V}_{h+1}^t(s)) - \rho_\beta^{p_h}(V_{h+1}^* - \mathring{V}_{h+1}^t(s)) &\leq \frac{1}{\beta}\log(1 + \frac{1}{H}\widehat{p}_h^t\big(e^{\beta V_{h+1}^*} - e^{\beta\mathring{V}_{h+1}^t}\big)(s,a) \\ &\quad + 8H^2 e^{\beta H}\frac{\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}) \\ &\leq \frac{1}{\beta}\log(1 + 8H^2 e^{\beta H}\frac{\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}) + \\ &\quad \frac{1}{\beta H}\widehat{p}_h^t\big(e^{\beta V_{h+1}^*} - e^{\beta\mathring{V}_{h+1}^t}\big)(s,a)\end{aligned}$$

Hence we find :

$$\begin{aligned}\rho_\beta^{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t(s)) - \rho_\beta^{p_h}(\mathring{V}_h^t(s)) &\leq (\rho_\beta^{\widehat{p}_h^t}(\widetilde{V}_{h+1}^t(s)) - \rho_\beta^{\widehat{p}_h^t}(\mathring{V}_{h+1}^t(s))) \\ &\quad + \log(1 + \widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)) + \frac{1}{H}\widehat{p}_t^h(e^{\beta\widetilde{V}_{h+1}^t} - e^{\beta V_{h+1}^t(s)}) \\ &\quad + \frac{1}{\beta}\log(1 + 8H^2 e^{\beta H}\frac{\beta\big(n_h^t(s,a),\delta\big)}{n_h^t(s,a)}) + \frac{1}{\beta H}\widehat{p}_h^t\big(e^{\beta V_{h+1}^*} - e^{\beta\mathring{V}_{h+1}^t}\big)(s,a)\end{aligned}$$

We conclude then by recurrence □

We derive then the sampling complexity :

Theorem 6.3 (PAC sample complexity for entropic BPI)**.** *Consider the algorithm greedy w.r.t. $\widetilde{Q}^t$ and the stopping time*

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

where G_h^t is the entropic recursion defined above Let the exploration rates be, for all $n \geq 1$,

$$\beta(n,\delta) = \log\frac{3SAH}{\delta} + S\log\big(8e(n+1)\big), \qquad \beta^\star(n,\delta) = \log\frac{3SAH}{\delta} + \log\big(8e(n+1)\big),$$

and set $\beta^{\mathrm{cnt}}(\delta) = \log\frac{3SAH}{\delta}$. Let $\mathcal{G} = \mathcal{E} \cap \mathcal{E}^{\mathrm{cnt}} \cap \mathcal{E}^\star$ be the high-probability event defined in the article; then $\Pr(\mathcal{G}) \geq 1 - \delta$.

Define the constants

$$L_\delta \triangleq \log\frac{3SAH}{\delta} + S, \qquad \Lambda_T \triangleq \log(8eT).$$

With probability at least $1-\delta$*, for every* $T<\tau$*,*

$$\sum_{t=0}^{T}\Big(V_1^\star(s_1)-V_1^{\pi^{t+1}}(s_1)\Big)\le \frac{288e^2e^{\beta H}}{\beta}\sqrt{T+1}\sqrt{H^3SA(L_\delta+\Lambda_T)}+\frac{288ee^{\beta H}}{\beta}H^3SA(L_\delta+\Lambda_T),$$

where the constants are explicitly *Consequently, we upper bound the sampling complexity :*

$$\boxed{\tau\le \frac{(252)^2e^{2\beta H}}{\beta^2}\frac{H^3SA}{\varepsilon^2}\Big(L_\delta+\overline{\Lambda}\Big)+\frac{(3024)e^{\beta H}}{\beta}\frac{H^3SA}{\varepsilon}\Big(L_\delta+\overline{\Lambda}\Big)+1}$$

where

$$\overline{\Lambda}=\log\left(\left[\frac{(252)^2e^{2\beta H}}{\beta^2}\frac{H^3SA}{\varepsilon^2}(L_\delta+1)+\frac{3024e^{\beta H}}{\beta}\frac{H^3SA}{\varepsilon}(L_\delta+1)+1\right]\right)$$

Proof. On the event $\mathcal{G}$ we have for all (t,h,s,a),

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

We prove then by recursion that :

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

Again on the event $\mathcal{E}^\star$ and using $\log(1+x)\le x$, for any (t,h,s) and $a=\pi_h^{t+1}(s)$,

$$\Big|\rho_\beta^{\widehat{p}_h^t}(V_{h+1}^\star)-\rho_\beta^{p_h}(V_{h+1}^\star)\Big|\le\frac{1}{\beta}\widetilde{s}_h^t(s,a)\widetilde{B}_h^t(s,a)\le\frac{1}{\beta}\widetilde{B}_h^t(s,a)$$

and the Bernstein-type bonus (applied to $f=e^{\beta V_{h+1}^\star}\in[1,e^{\beta H}]$) gives

$$G_h^t(s)\le\min\Big\{H,\frac{6}{\beta}\sqrt{\mathrm{Var}_{\widehat{p}_h^t}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a)\cdot\frac{\beta^\star(n_h^t,\delta)}{n_h^t}}+\frac{36e^{\beta H}}{\beta}\cdot\frac{\beta(n_h^t,\delta)}{n_h^t}+\Big(1+\frac{3}{H}\Big)\big(\widehat{p}_h^t\pi_{h+1}^{t+1}G_{h+1}^t\big)(s,a)\Big\}\tag{2}$$

Unrolling the previous inequality along the episode trajectory under π^{t+1} and using $\prod_{j=1}^{h-1}(1+\frac{3}{H})\le(1+\frac{3}{H})^H\le e^3<21$, we get, for each episode t,

$$\begin{aligned}\pi_1^{t+1}G_1^t(s_1)&\le\frac{126}{\beta}\sum_{h,s,a}\widehat{p}_h^{t,\pi}(s,a)\sqrt{\mathrm{Var}_{\widehat{p}_h^t}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a)\cdot\frac{\beta^\star(n_h^t,\delta)}{n_h^t}}\\&\quad+\frac{756e^{\beta H}}{\beta}\sum_{h,s,a}\widehat{p}_h^{t,\pi}(s,a)\ \frac{\beta(n_h^t,\delta)}{n_h^t}\qquad\text{(define these as } Y_t \text{ and } W_t).\end{aligned}$$

For the Y_t term, we apply Cauchy-Schwartz inequality :

$$Y_t\ \le\frac{126}{\beta}\sqrt{\sum_{h,s,a}\widehat{p}_h^{t,\pi}(s,a)\mathrm{Var}_{\widehat{p}_h^t}\big(e^{\beta\widetilde{V}_{h+1}^t}\big)(s,a)}\cdot\sqrt{\sum_{h,s,a}\widehat{p}_h^{t,\pi}(s,a)\frac{\beta^\star(n_h^t,\delta)}{n_h^t}}$$

Since $e^{\beta\widetilde{V}^t_{h+1}} \in [1, e^{\beta H}]$ we use $\mathrm{Var}(e^{\beta\widetilde{V}^t_{h+1}}) \le (e^{\beta H}-1)^2/4$ and the Bellman-variance telescoping (applied to the occupancy $\widehat{p}^{t,\pi}_h$) to obtain

$$\sum_{h,s,a} \widehat{p}^{t,\pi}_h(s,a)\mathrm{Var}_{\widehat{p}^t_h}\big(e^{\beta\widetilde{V}^t_{h+1}}\big)(s,a) \le e^{\beta H}-1)^2/4H$$

Hence

$$Y_t \;\le\; \frac{126e^{\beta H}}{\beta}\sqrt{H}\sqrt{\sum_{h,s,a}\widehat{p}^{t,\pi}_h(s,a)\frac{\beta^\star(n^t_h,\delta)}{n^t_h}}$$

Summing over $t = 0, 1, \dots, T$ and using Cauchy–Schwarz across episodes gives

$$\sum_{t=0}^{T}\pi_1^{t+1}G^t_1(s_1) \le \frac{126e^{\beta H}}{\beta}\sqrt{H}\sqrt{T+1}\sqrt{\sum_{t,h,s,a}\widehat{p}^{t,\pi}_h(s,a)\frac{\beta^\star(n^t_h,\delta)}{n^t_h}}+\frac{756e^{\beta H}}{\beta}\sum_{t,h,s,a}\widehat{p}^{t,\pi}_h(s,a)\frac{\beta(n^t_h,\delta)}{n^t_h}$$

We use the counting inequalities, for all $T \ge 1$:

$$\sum_{t,h,s,a}\widehat{p}^{t,\pi}_h(s,a)\frac{\beta^\star(n^t_h,\delta)}{n^t_h} \le 4H^3SA\big(L_\delta+\Lambda_T\big) \qquad \sum_{t,h,s,a}\widehat{p}^{t,\pi}_h(s,a)\frac{\beta(n^t_h,\delta)}{n^t_h} \le 4H^3SA\big(L_\delta+\Lambda_T\big)$$

Substituting into the inequality yields

$$\sum_{t=0}^{T}\pi_1^{t+1}G^t_1(s_1) \;\le\; \underbrace{\frac{126(e^{\beta H}-1)}{\beta}\sqrt{H}\sqrt{4}}_{=\frac{252(e^{\beta H}-1)}{\beta}\sqrt{H}}\sqrt{T+1}\sqrt{H^3SA(L_\delta+\Lambda_T)}+\underbrace{\frac{756e^{\beta H}}{\beta}\cdot 4}_{=\frac{3024e^{\beta H}}{\beta}}H^3SA(L_\delta+\Lambda_T)$$

Hence, by using the definition of τ for all $T < \tau$,

$$(T+1)\varepsilon \le \frac{252(e^{\beta H}-1)}{\beta}\sqrt{T+1}\sqrt{H^3SA(L_\delta+\Lambda_T)}+\frac{3024e^{\beta H}}{\beta}H^3SA(L_\delta+\Lambda_T),$$

We solve this inequality using lemma 13 from (Ménard et al., 2021) and we complete the proof. □

6.3 Proofs of section 5

6.3.1 Bernstein inequality

Let us try to establish a Bernstein style bound in the curvature Let $p, q \in \Sigma_S$ and $f : \mathcal{S} \to \mathbb{R}$ such that $f \in [0, b]$. Fix $\beta \in \mathbb{R}$ Remark that if we write $w = \frac{dp}{dq}$ then :

$$pe^{\beta h} = qwe^{\beta h} \le (qw^{1+\lambda})^{\frac{1}{1+\lambda}}(qe^{\frac{1+\lambda}{\lambda}\beta h})^{\frac{\lambda}{1+\lambda}}$$

Which gives by taking the logarithms that :

$$\log p[e^{\beta h}] - \log q[e^{\beta h}](\beta) \leq \frac{1}{1+\lambda}\log q[w^{1+\lambda}] + \frac{\lambda}{1+\lambda}\Big(A_q\big(\tfrac{1+\lambda}{\lambda}\beta\big) - A_q(\beta)\Big)$$

Recognizing $\log \mathbb{E}_q[w^{1+\lambda}] = D_{1+\lambda}(p\|q)$, we obtain

$$\log \mathbb{E}_p[e^{\beta g}] - A_q(\beta) \leq \frac{\lambda}{1+\lambda} D_{1+\lambda}(p\|q) + \frac{\lambda}{1+\lambda}\Big(A_q\big(\tfrac{1+\lambda}{\lambda}\beta\big) - A_q(\beta)\Big)$$

Now remark that we can write :

$$q_\beta e^{\frac{\beta}{\lambda}g} = \frac{q e^{\frac{\beta}{\lambda}} e^{\beta g}}{q e^{\beta g}}$$

Now let's take $h = g - q_\beta g$

Lemma 6.6 (Bernstein inequality)**.**

Proof. Let $p, q \in \Sigma_S$ and $f : \mathcal{S} \to \mathbb{R}$ such that $f \in [0, \beta]$. Fix $\beta \in \mathbb{R}$. We define the tilted exponential distribution as :

$$q_\beta(ds) = \frac{e^{\beta f(s)}}{q e^{\beta f}} q(ds)$$

Now remark that the if we denote $g(s) = f(s) - q_\beta f$, if we considered the tilted distribution defined from g q_β^g then :

$$q_\beta^g(ds) = \frac{e^{\beta(f(s) - q_\beta f)}}{q e^{\beta(f - q_\beta f)}} q(ds) = q_\beta(s)$$

And q_b is shift invariant with respect to f. Now, denote $w = \frac{dp}{dq}$ then :

$$p e^{\beta g} = q w e^{\beta g} \leq (q w^{1+\lambda})^{\frac{1}{1+\lambda}} (q e^{\frac{1+\lambda}{\lambda}\beta g})^{\frac{\lambda}{1+\lambda}}$$

We take the logarithm :

$$A_p^{(g)}(\beta) \leq \frac{1}{1+\lambda}\log q[w^{1+\lambda}] + \frac{\lambda}{1+\lambda} A_q^{(g)}\big(\tfrac{1+\lambda}{\lambda}\beta\big)$$

Where we denoted $A_q(\beta) = \log(q e^{\beta f})$ and $A_q^{(g)}(\beta) = \log(q e^{\beta g})$. Now notice that we can write :

$$q_\beta e^{\frac{\beta}{\lambda}g} = q_\beta^{(g)} e^{\frac{\beta}{\lambda}g} = \frac{q e^{\frac{\beta}{\lambda}g} e^{\beta g}}{q e^{\beta g}}$$

And by taking the log :

$$A_q^{(g)}\big(\tfrac{1+\lambda}{\lambda}\beta\big) = \log(q_\beta e^{\frac{\beta}{\lambda}g}) + A_q^{(g)}(\beta)$$

Now, notice that :

$$\log(q e^{\beta g}) = \log(q e^{\beta f} e^{-\beta q_\beta f}) = A_q(\beta) - \beta q_\beta f$$

On the other hand :

$$D_{KL}(q_\beta||q) = q_\beta \log(\tfrac{dq_\beta}{dq}) = q_\beta \log(\tfrac{e^{\beta f}}{qe^{\beta f}}) = \beta q_\beta f - A_q(\beta) = -A_q^{(g)}(\beta)$$

Hence

$$A_q^{(g)}\big(\tfrac{1+\lambda}{\lambda}\beta\big) \leq \log(q_\beta e^{\frac{\beta}{\lambda}g})$$

Hence :

$$A_p(\beta) - A_q(\beta) \leq \frac{1}{1+\lambda}\log q[w^{1+\lambda}] + \frac{\lambda}{1+\lambda}\log(q_\beta e^{\frac{\beta}{\lambda}g})$$

Now recognize the Reyni divergence :

$$D_{1+\lambda}(p||q) = \frac{1}{\lambda}\log q[w^{1+\lambda}]$$

Hence :

$$A_p(\beta) - A_q(\beta) \leq \frac{\lambda}{1+\lambda}D_{1+\lambda}(p||q) + \frac{\lambda}{1+\lambda}\log(q_\beta e^{\frac{\beta}{\lambda}g})$$

Now consider the lemma :

Lemma 6.7. *Let X be a mean–zero random variable with $|X| \leq R$ almost surely and $\mathbb{V}ar(X) = \sigma^2$ then for all $t \in \mathbb{R}$,*

$$\log(\mathbb{E}[e^{tX}]) \leq \frac{\sigma^2}{R^2}\phi(|t|R)$$

Lemma. For $u \geq 0$ and $|y| \leq R$ expand and bound the exponential series

$$e^{ty} - 1 - ty = \sum_{k=2}^{\infty}\frac{(ty)^k}{k!} = y^2\sum_{k=0}^{\infty}\frac{t^{k+2}y^k}{(k+2)!} \leq y^2\sum_{k=0}^{\infty}\frac{t^{k+2}R^k}{(k+2)!} = \frac{y^2}{2}\phi(tR)$$

Where $\phi(z) = e^z - z - 1$. Take the expectation in the previous inequality gives :

$$\mathbb{E}[e^{tX}] \leq 1 + \frac{\sigma^2}{R^2}\phi(tR)$$

The same holds for $t \leq 0$ with $|t|$ and we find the inequality :

$$\mathbb{E}[e^{tX}] \leq 1 + \frac{\sigma^2}{R^2}\phi(|t|R)$$

Taking the logarithm and using the inequality $\log(1+x) \leq x$ yields :

$$\log(\mathbb{E}[e^{tX}]) \leq \frac{\sigma^2}{R^2}\phi(|t|R)$$

□

We apply this lemma to $\log(q_\beta e^{\frac{\beta}{\lambda}g})$ and we have the bound :

$$A_p(\beta) - A_q(\beta) \leq \frac{\lambda}{1+\lambda} D_{1+\lambda}(p||q) + \frac{\lambda}{1+\lambda} \frac{\mathrm{Var}_{q_\beta}(g)}{R^2} \phi(\beta H)$$

And dividing by $|\beta|$ and using that $\mathrm{Var}_{q_\beta}(f) = \mathrm{Var}_{q_\beta}(g)$:

$$sgn(\beta)(\rho_{\beta,p}(f) - \rho_{\beta,q}(g)) \leq \frac{\lambda}{|\beta|(1+\lambda)} D_{1+\lambda}(p||q) + \frac{\lambda}{|\beta|(1+\lambda)} \frac{\mathrm{Var}_{q_\beta}(f)}{R^2} \phi(\frac{|\beta| H}{\lambda})$$

□

6.3.2 Reyni Divergence and impossibility of a concentration result

We define the Reyni divergence as : Now we would love to have concentration bounds on Reyni divergence just like the KL divergence case which would allows us to derive bonus terms. Unfortunately, this is not always possible

Theorem 6.4. *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*

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

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

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

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 $b(n,\delta) = O(\log(1/\delta))$

Proof. Let $\alpha > 1$. Choose a constant

$$c \in \left(0,\ \min\{1/2,\ (4\alpha)^{-1/(\alpha-1)}\}\right]$$

and then pick $0 < \varepsilon \leq \min\{c/2,\ c^\alpha/2\}$. Define the fixed binary distribution

$$p = (\varepsilon,\ 1-\varepsilon).$$

And chose the time

$$n_\star := \lfloor c/\varepsilon \rfloor\ (\geq 2)$$

Consider the event

$$A := \{N_1 = 1\}, \qquad N_1 \sim \mathrm{Binom}(n_\star, \varepsilon)$$

i.e., among the first $n_\star$ draws, the rare category appears exactly once. Conditionally on A the empirical law is

$$\hat{p}_{n_\star} = \left(\tfrac{1}{n_\star},\ 1 - \tfrac{1}{n_\star}\right)$$

The probability of A is given by:

$$\Pr_p(A) = n_\star \varepsilon (1-\varepsilon)^{n_\star - 1}$$

Since $n_\star \geq c/\varepsilon - 1$ and $\varepsilon \leq c/2$,

$$n_\star \varepsilon \;\geq\; c - \varepsilon \;\geq\; c/2$$

Using $\log(1-x) \geq -x/(1-x)$ and $\varepsilon \leq 1/2$,

$$(1-\varepsilon)^{n_\star - 1} \;\geq\; \exp\left(-\frac{(n_\star - 1)\varepsilon}{1-\varepsilon}\right) \;\geq\; e^{-2c}$$

Hence

$$\Pr_p(A) \;\geq\; \tfrac{c}{2} e^{-2c}$$

On the other hand:

$$D_\alpha(\hat{p}\|p) = \frac{1}{\alpha - 1} \log \Big[\underbrace{\left(\tfrac{1}{n_\star}\right)^\alpha \varepsilon^{1-\alpha}}_{T_1} + \underbrace{\left(1 - \tfrac{1}{n_\star}\right)^\alpha (1-\varepsilon)^{1-\alpha}}_{T_2} \Big]$$

Since $1 - \alpha < 0$ and $\varepsilon \leq 1/2$, $(1-\varepsilon)^{1-\alpha} \geq 1$, and by Bernoulli, $(1 - \frac{1}{n_\star})^\alpha \geq 1 - \frac{\alpha}{n_\star}$. Thus

$$D_\alpha(\hat{p}\|p) \;\geq\; \frac{1}{\alpha - 1} \log\left(1 - \tfrac{\alpha}{n_\star} + T_1\right)$$

Since $n_\star \leq c/\varepsilon$,

$$T_1 \;\geq\; \varepsilon^{1-\alpha} \left(\tfrac{\varepsilon}{c}\right)^\alpha = \varepsilon\, c^{-\alpha}.$$

Define $y := T_1 - \frac{\alpha}{n_\star}$. Because $n_\star \geq c/\varepsilon - 1$,

$$n_\star y = n_\star T_1 - \alpha \;\geq\; \left(\tfrac{c}{\varepsilon} - 1\right) \varepsilon c^{-\alpha} - \alpha = c^{1-\alpha} - \varepsilon c^{-\alpha} - \alpha.$$

With $\varepsilon \leq c/2$ and $c^{1-\alpha} \geq 4\alpha$ (by our choice of c),

$$n_\star y \;\geq\; \tfrac{1}{4}\, c^{1-\alpha}$$

Also $y \leq T_1 \leq \varepsilon c^{-\alpha} \leq \frac{1}{2}$ (by $\varepsilon \leq c^\alpha/2$), hence

$$y \in [0, 1/2]$$

Using $\log(1+u) \geq u/2$ for $u \in [0,1]$ and $1 - \frac{\alpha}{n_\star} + T_1 = 1 + y$,

$$D_\alpha(\hat{p}\|p) \;\geq\; \frac{y}{2(\alpha - 1)}$$

Multiplying by $n_\star$ we get,

$$n_\star D_\alpha(\hat{p}_{n_\star}\|p) \;\geq\; \frac{1}{8(\alpha-1)}\, c^{1-\alpha}$$

Now, let us chose c as 4δ. For δ small enough (e.g., $\delta \leq \delta_0 := \min\{1/8,\ (4\alpha)^{-1/(\alpha-1)}/4\}$) all conditions hold

$$\Pr_p(A) \;\geq\; \tfrac{c}{2} e^{-2c} \;=\; 2\delta\, e^{-8\delta} \;\geq\; \delta.$$

And

$$n_\star D_\alpha(\hat{p}_{n_\star}\|p) \;\geq\; \frac{1}{8(\alpha-1)}\, c^{1-\alpha} = \;\frac{4^{\,1-\alpha}}{8(\alpha-1)}\delta^{-(\alpha-1)}$$

Let $t := \frac{4^{\,1-\alpha}}{8(\alpha-1)}\delta^{-(\alpha-1)}$. Then on A we have $n_\star D_\alpha \geq t$, hence

$$\Pr_p\{\exists n \geq 1:\ nD_\alpha(\hat{p}_n\|p) \geq t\} \;\geq\; \Pr_p\{n_\star D_\alpha(\hat{p}_{n_\star}\|p) \geq t\} \;\geq\; \Pr_p(A) \;\geq\; \delta$$

If we assume that there is a an upper bound bound $b(n,\delta)$ in $\log(\frac{1}{\delta})$ that holds uniformly. For δ small enough $b(n,\delta) \leq t$ Hence

$$\Pr_p\{\exists n \geq 1:\ nD_\alpha(\hat{p}_n\|p) \geq b(n,\delta)\} \geq \Pr_p\{\exists n \geq 1:\ nD_\alpha(\hat{p}_n\|p) \geq t\} \geq \delta$$

Contradicting that $b(n,\delta)$ is an upper bound uniformly in n with probability $1-\delta$ □

Let us try to establish bounds on Renyi divergence after having a bounded like hood.

Theorem 6.5 (All-time Renyi bound after Chernoff burn-in)**.** *Let $p \in \Delta S$ let $\hat{p}_n$ be its statistical estimation, and suppose $p_{\min} \doteq \min_i p_i \geq b > 0$. 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

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

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

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

Proof. We intersect two high-probability events that both hold uniformly in n. For each category i, $\hat{p}_{n,i} = \frac{1}{n}\sum_{t=1}^n \mathbf{1}\{X_t = i\}$ satisfies, for $0 < \eta \leq 1$ (two-sided multiplicative Chernoff),

$$\Pr\big(|\hat{p}_{n,i} - p_i| > \eta p_i\big) \leq 2\exp\Big(-\tfrac{\eta^2}{3}np_i\Big) \leq 2\exp\Big(-\tfrac{\eta^2}{3}nb\Big).$$

Union over $i \in [S]$ and over all $n \geq n_0$ (sum a geometric tail) yields

$$\Pr\left(\exists n \geq n_0, \exists i : \ |\hat{p}_{n,i} - p_i| > \eta p_i\right) \leq \frac{2Se^{-\eta^2 b n_0/3}}{1 - e^{-\eta^2 b/3}}.$$

Choose $\eta = \frac{1}{2}$ and n_0 (using $1 - e^{-x} \geq x/2$ for $x \in (0, 1]$) to make the RHS $\leq \delta/2$. Hence, with probability $\geq 1 - \delta/2$, the event

$$\mathcal{E}_{\mathrm{LR}} : \qquad \forall n \geq n_0, \ \forall i : \quad \tfrac{1}{2} p_i \leq \hat{p}_{n,i} \leq \tfrac{3}{2} p_i$$

holds, i.e. the likelihood ratios $r_i \doteq \hat{p}_{n,i}/p_i$ satisfy $r_i \in [1/2, 3/2]$ for every $n \geq n_0$.

Now, Write $\chi(\hat{p}\|p) = \sum_i p_i (r_i - 1)^2$ and $D_{\mathrm{KL}}(\hat{p}\|p) = \sum_i p_i (r_i \log r_i - r_i + 1)$. Under the upper cap $\max_i r_i \leq L_0$, the standard f-divergence comparison (via the second-derivative ratio $f''_{\chi^2}(t)/f''_{\mathrm{KL}}(t) = 2t \leq 2L_0$ and Taylor's theorem with integral remainder) gives

$$\chi(\hat{p}\|p) \ \leq \ 2L_0 D_{\mathrm{KL}}(\hat{p}\|p)$$

For $1 < \alpha \leq 2$, by monotonicity $D_\alpha \leq D_2 = \log(1 + \chi) \leq \chi$, hence on $\mathcal{E}_{\mathrm{LR}}$,

$$D_\alpha(\hat{p}\|p) \ \leq \ \chi(\hat{p}\|p) \ \leq \ 2L_0 D_{\mathrm{KL}}(\hat{p}\|p)$$

For $\alpha > 2$, for $t \in (0, L_0]$ the Taylor bound $t^\alpha \leq 1 + \alpha(t - 1) + \frac{\alpha(\alpha-1)}{2} L_0^{\alpha-2} (t - 1)^2$ holds. Averaging under p (the linear term cancels because $\sum_i p_i (r_i - 1) = 0$) gives

$$\sum_i p_i r_i^\alpha \ \leq \ 1 + \tfrac{\alpha(\alpha-1)}{2} L_0^{\alpha-2} \chi(\hat{p}\|p)$$

so using $\log(1 + x) \leq x$

$$D_\alpha(\hat{p}\|p) \ = \ \tfrac{1}{\alpha-1} \log \sum_i p_i r_i^\alpha \leq \tfrac{\alpha}{2} L_0^{\alpha-2} \chi(\hat{p}\|p) \leq \alpha L_0^{\alpha-1} D_{\mathrm{KL}}(\hat{p}\|p)$$

We then combine the two events to get the inequality with probability $1 - \delta$ □

References

Amir Ahmadi-Javid. Entropic value-at-risk: A new coherent risk measure. *Journal of Optimization Theory and Applications*, 155(3):1105–1123, 2012. doi: 10.1007/s10957-011-9968-2.

Kenneth J. Arrow. *Essays in the Theory of Risk-Bearing*. North-Holland, Amsterdam, 1971. ISBN 978-0720430479.

Philippe Artzner, Freddy Delbaen, Jean-Marc Eber, and David Heath. Coherent measures of risk. *Mathematical Finance*, 9(3):203–228, 1999. doi: 10.1111/1467-9965.00068. URL `https://doi.org/10.1111/1467-9965.00068`.

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.

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

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

Vivek S. Borkar and Sean P. Meyn. The o.d.e. method for convergence of stochastic approximation and reinforcement learning. *SIAM Journal on Control and Optimization*, 38(2): 447–469, 2000. doi: 10.1137/S0363012997331639.

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.

Kai Detlefsen and Giacomo Scandolo. Conditional and dynamic convex risk measures. *Finance and Stochastics*, 9(4):539–561, 2005. doi: 10.1007/s00780-005-0151-1.

Yingjie Fei, Zhuoran Yang, Yudong Chen, Zhaoran Wang, and Qiaomin Xie. Risk-sensitive reinforcement learning: Near-optimal risk-sample tradeoff in regret. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL `https://proceedings.neurips.cc/paper/2020/hash/fdc42b6b0ee16a2f866281508ef56730-Abstract.html`.

Yingjie Fei, Zhuoran Yang, and Zhaoran Wang. Risk-sensitive reinforcement learning with function approximation: A debiasing approach. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 3198–3207. PMLR, 2021. URL `https://proceedings.mlr.press/v139/fei21a.html`.

Hans Föllmer and Alexander Schied. Convex measures of risk. *Finance and Stochastics*, 6 (2):173–197, 2002. doi: 10.1007/s007800200072.

Alonso Granados, Mohammadreza Ebrahimi, and Jason Pacheco. Risk-sensitive variational actor-critic: A model-based approach. In *International Conference on Learning Representations (ICLR)*, 2025. URL `https://openreview.net/forum?id=irrtPRFksw`.

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.

David H. Jacobson. Optimal stochastic linear systems with exponential performance criteria and their relation to deterministic differential games. *IEEE Transactions on Automatic Control*, 18(2):124–131, 1973. doi: 10.1109/TAC.1973.1100265.

Thomas Jaksch, Ronald Ortner, and Peter Auer. Near-optimal regret bounds for rein-

forcement learning. *Journal of Machine Learning Research*, 11:1563–1600, 2010. URL https://jmlr.org/papers/v11/jaksch10a.html.

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.

Emilie Kaufmann, Olivier Cappé, and Aurélien Garivier. On the complexity of best arm identification in multi-armed bandit models. *Journal of Machine Learning Research*, 17 (1):1–42, 2016. URL https://jmlr.org/papers/v17/. Includes the change-of-measure (transportation) lemma used for BAI.

Emilie Kaufmann, Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Edouard Leurent, and Michal Valko. Adaptive reward-free exploration. In *Proceedings of the 32nd International Conference on Algorithmic Learning Theory (ALT)*, volume 132 of *Proceedings of Machine Learning Research*, pages 865–891. PMLR, 2021. URL https://proceedings.mlr.press/v132/kaufmann21a.html.

Alexandre Marthe, Aurélien Garivier, and Claire Vernade. Beyond average return in markov decision processes. *Advances in Neural Information Processing Systems*, 36:56488–56507, 2023.

Alexandre Marthe, Samuel Bounan, Aurélien Garivier, and Claire Vernade. Efficient risk-sensitive planning via entropic risk measures. *arXiv preprint*, 2025. doi: 10.48550/arXiv.2502.20423. URL https://arxiv.org/abs/2502.20423.

Pierre Ménard, Omar Darwiche Domingues, Anders Jonsson, Emilie Kaufmann, Edouard Leurent, and Michal Valko. Fast active learning for pure exploration in reinforcement learning. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, volume 139 of *Proceedings of Machine Learning Research*, pages 7599–7608. PMLR, 2021. URL https://proceedings.mlr.press/v139/menard21a.html.

Oliver Mortensen and Mohammad Sadegh Talebi. Entropic risk optimization in discounted mdps: Sample complexity bounds with a generative model. *arXiv preprint arXiv:2506.00286*, 2025. URL https://arxiv.org/abs/2506.00286.

John W. Pratt. Risk aversion in the small and in the large. *Econometrica*, 32(1/2):122–136, 1964. doi: 10.2307/1913738.

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. Wiley Series in Probability and Mathematical Statistics. John Wiley & Sons, New York, 1994. ISBN 978-0-471-61977-2. doi: 10.1002/9780470316887. See Theorem 4.4.2.

R. Tyrrell Rockafellar and Stanislav Uryasev. Optimization of conditional value-at-risk. *Journal of Risk*, 2(3):21–41, 2000.

Andrzej Ruszczyński. Risk-averse dynamic programming for markov decision processes. *Mathematical Programming*, 125(2):235–261, 2010. doi: 10.1007/s10107-010-0393-3.

Aviv Tamar, Yinlam Chow, Mohammad Ghavamzadeh, and Shie Mannor. Policy gradient for coherent risk measures. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015. URL `https://papers.nips.cc/paper/5923-policy-gradient-for-coherent-risk-measures`.

Peter Whittle. *Risk-Sensitive Optimal Control*. John Wiley & Sons, Chichester, 1990. ISBN 978-0-471-92622-1.

Minheng Xiao, Xian Yu, and Lei Ying. Policy gradient methods for risk-sensitive distributional reinforcement learning with provable convergence. *arXiv preprint arXiv:2405.14749*, 2024. URL `https://arxiv.org/abs/2405.14749`.

Andrea Zanette and Emma Brunskill. Tighter problem-dependent regret bounds in reinforcement learning without domain knowledge using value function bounds. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7304–7312. PMLR, 2019. URL `https://proceedings.mlr.press/v97/zanette19a.html`.

Bernardo Ávila Pires, Mark Rowland, Diana Borsa, Zhaohan Daniel Guo, Khimya Khetarpal, André Barreto, David Abel, Rémi Munos, and Will Dabney. Optimizing return distributions with distributional dynamic programming. *Journal of Machine Learning Research*, 26(185):1–90, 2025.