Distributional Policy Evaluation: a Maximum Entropy approach to Representation Learning

Anonymous Author(s)
Affiliation
Address
email

Abstract

The Maximum Entropy (Max-Ent) framework has been effectively employed in a variety of Reinforcement Learning (RL) tasks. In this paper, we first propose a novel Max-Ent framework for policy evaluation in a distributional RL setting, named *Distributional Maximum Entropy Policy Evaluation* (D-Max-Ent PE). We derive a generalization-error bound that depends on the complexity of the representation employed, showing that this framework can explicitly take into account the features used to represent the state space while evaluating a policy. Then, we exploit these favorable properties to drive the representation learning of the state space in a Structural Risk Minimization fashion. We employ state-aggregation functions as feature functions and we specialize the D-Max-Ent approach into an algorithm, named *D-Max-Ent Progressive Factorization*, which constructs a progressively finer-grained representation of the state space by balancing the trade-off between preserving information (bias) and reducing the effective number of states, i.e., the complexity of the representation space (variance). Finally, we report the results of some illustrative numerical simulations, showing that the proposed algorithm matches the expected theoretical behavior and highlighting the relationship between aggregations and sample regimes.

1 Introduction

In Distributional Reinforcement Learning (D-RL) [Bellemare et al., 2023], an agent aims to estimate the entire distribution of the returns achievable by acting according to a specific policy. This is in contrast to and more complex than classic Reinforcement Learning (RL) [Szepesvári, 2010, Sutton and Barto, 2018], where the objective is to predict the expected return only. In recent years, several algorithms for D-RL have been proposed, both in evaluation and control settings. The push towards distributional approaches was particularly driven by additional flavors they can bring into the discourse, such as risk-averse considerations, robust control, and many regularization techniques [Chow et al., 2015, Brown et al., 2020, Keramati et al., 2020]. Most of them varied in how the distribution of the returns is modeled. The choice of the model was shown to have a cascading effect on how such a distribution can be learned, how efficiently and with what guarantees, and how it can be used for the control problem. Similarly to this tradition, this paper investigates the potential of looking into the entire distribution of returns to address the representation learning of the state-action spaces. In particular, it points to answer the following research question:

> Q1: *Does a distributional approach offer novel tools to address the representation learning problem?*

In Section 3, we answer this methodological question, showing that it is possible to reformulate Policy Evaluation in a distributional setting so that its performance index is explicitly intertwined

Submitted to 16th European Workshop on Reinforcement Learning (EWRL 2023). Do not distribute.

with the representation of the (state or action) spaces. More specifically, this work tackles Policy Evaluation in a distributional setting as a particular case of a distribution estimation problem and then applies the Maximum Entropy (Max-Ent) formulation of distribution estimation [Wainwright and Jordan, 2007]. In this way, it is possible to derive a novel framework for PE, which we name *Distributional Max-Ent Policy Evaluation* (D-Max-Ent PE), which inherits the many positives the Max-Ent framework offers. In particular, it allows the inclusion of constraints that the distribution needs to satisfy, usually called *structural constraints*, namely feature-based constraints acting over the support of the distribution. Such constraints then appear in the generalization-error bound for the Max-Ent problem, with a term related to the complexity of the family of features used. Unfortunately, traditional derivations of this bound introduce quantities that are bounded but unknown. We develop the analysis further to derive a more practical bound, containing quantities that are either estimated or known. In this way, the generalization-error bound shows a usual bias-variance trade-off that can be explicitly optimized by changing the feature functions adopted, while making the best use of the available samples. Thus, PE in a distributional setting is directly linked to representation learning.

Now, the RL literature proved that reducing the state space size while preserving the important features of the original state space is beneficial, namely with state-aggregation feature functions [Singh et al., 1994, Van Roy, 2006, Dong et al., 2020]. This is particularly true when high dimensionality can make learning slower and more unstable, as in classic RL in general, or when the learning process is almost unfeasible in small-samples regimes, as for D-RL, where learning the entire distribution of returns requires a large number of samples. Thus, motivated by these considerations, while D-Max-Ent Policy Evaluation allows for the use of any type of structural constraint, this work focuses on state-aggregation feature functions, and we exploit the first and more methodological result to answer a second more algorithmic question:

> Q2: *Is it beneficial to drive the learning of a refined state-aggregated representation based on the evaluation of the quality of a policy in a distributional setting?*

To answer this question, in Section 4, we show that the generalization-error bound changes monotonically when the state-aggregation constraints are changed in a specific way, namely a finer-grained representation of the space. This is indeed what happens in the Structural Risk Minimization theory (SRM) [Vapnik, 1991]. Similarly, we develop a novel algorithm called *D-Max-Ent Progressive Factorization*, which exploits the proposed evaluation method to learn a representation of the state space soundly, i.e., trying to reduce a proxy of the generalization error bound in an SRM fashion, and allowing us to answer positively to the second research question as well. Finally, in Section 5 we verify through an illustrative numerical simulation whether the proposed algorithm matches the behaviors suggested by the theoretical analysis.

2 Preliminaries

2.1 Markov Decision Processes

A discrete–time finite Markov decision processes (MDP) [Puterman, 1994] is a tuple $\mathcal{M} := (\mathcal{S}, \mathcal{A}, P_{\mathcal{S}}, P_{\mathcal{R}}, \mu, \gamma)$, where $\mathcal{S}$ is a finite state space ($|\mathcal{S}| = S$), $\mathcal{A}$ is a finite action space ($|\mathcal{A}| = A$), $P_{\mathcal{S}} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathcal{S})$ is the transition kernel, $P_{\mathcal{R}} : \mathcal{S} \times \mathcal{A} \to \Delta(\mathbb{R})$ is the reward distribution function, $\mu \in \Delta(\mathcal{S})$ is the initial-state distribution and $\gamma \in [0, 1)$ is discount factor.[1] A policy $\pi : \mathcal{S} \to \Delta(\mathcal{A})$ defines the behavior of an agent interacting with an environment, which goes as follows: starting from an initial state $S_0 \sim \mu$, an agent interacts with the environment through the policy π, generating a trajectory $\mathcal{H} = (S_t, A_t, R_t)_{t=0}^{\infty}$, which is a sequence of states, actions and rewards whose joint distribution is determined by the transition kernel, reward distribution, and the policy itself, i.e., $A_t \sim \pi(\cdot|S_t)$, $R_t \sim P_{\mathcal{R}}(\cdot|S_t, A_t)$, and $S_{t+1} \sim P_{\mathcal{S}}(\cdot|S_t, A_t)$.

2.2 Value Functions and Distributions of Returns

Given an MDP $\mathcal{M}$ with discount factor γ, the *Discounted Return* is the sum of rewards received from the initial state onwards, discounted according to their time of occurrence:

[1] $\Delta(\mathcal{X})$ denotes the simplex of a space $\mathcal{X}$.

$$\mathcal{G}^\pi(s) = \sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s. \tag{1}$$

The *Value Function* of a given policy π is the expectation of this quantity under the policy itself:

$$V^\pi(s) = \mathbb{E}[\mathcal{G}(s)] = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s\right]. \tag{2}$$

The *Return Distribution Function* η^π of a given policy π is a collection of distributions, one for each state $s \in \mathcal{S}$, where each element is the distribution of the random variable $\mathcal{G}^\pi(s)$:

$$\eta^\pi(s) = \mathcal{D}^\pi_{(s)}\left[\sum_{t=0}^{\infty} \gamma^t R_t | S_0 = s\right], \tag{3}$$

where $\mathcal{D}^\pi_{(s)}$ extracts the probability distribution of a random variable under the joint distribution of the trajectory.

2.3 Maximum Entropy Estimation

Maximum Entropy (Max-Ent) methods [Dudík and Schapire, 2006, Wainwright and Jordan, 2007, Sutter et al., 2017] are density estimation methods that select the distribution that maximizes the uncertainty, i.e., the one with maximum entropy, where the entropy of a distribution $p \in \Delta(\mathcal{X})$ is defined as $H(p) \coloneqq -\mathbb{E}_{X\sim p}[\log p(X)]$.[2] Additionally, they assume that the learner has access to a feature mapping $\mathcal{F}$ from $\mathcal{X}$ to $\mathbb{R}^M$. In the most general case, we may have $M = +\infty$. We will denote by Φ the class of real-valued functions containing the component feature functions $f_j \in \mathcal{F}$ with $j \in [M]$. A distribution p is *consistent* with the true underlying distribution p_0 if

$$\mathbb{E}_{X\sim p}[f_j(X)] = \mu_j, \quad \forall j \in [M], \tag{4}$$

where

$$\mu_j \coloneqq \mathbb{E}_{X\sim p_0}[f_j(X)] \tag{5}$$

In this case, we say that p satisfies (in expectation) the structural constraints imposed by the features in $\mathcal{F}$. In practice, p_0 is not available and Max-Ent methods enforce empirical consistency over N independent and i.i.d. observations $\mathcal{D} = \{x_1, \ldots, x_N\} \sim p_0$ with support in $\mathcal{X}$ by replacing the definition in Eq. (5) with

$$\hat{\mu}_j(\mathcal{D}) \coloneqq \frac{1}{N}\sum_{i=1}^{N} f_j(x_i), \quad \forall j \in [M]. \tag{6}$$

The distribution p is said to be consistent with the data $\mathcal{D}$ if it matches the empirical expectations. The empirical Max-Ent problem consists then of the following optimization problem

$$\begin{aligned} \max_{p\in\Delta(\mathcal{X})} \quad & H(p) \\ \text{s.t.} \quad & \mathbb{E}_{X\sim p}[f_j(X)] = \hat{\mu}_j, \quad \forall j \in [M], \end{aligned} \tag{7}$$

with the optimization problem in expectation differing just in the constraints (i.e., replacing constraint from Eq. (6) with the ones from Eq. (5)). It is well known that the optimal solution to the empirical Max-Ent problem in Eq. (7) is a distribution $p_\lambda \in \Delta(\mathcal{X})$ belonging to the class of exponential distributions parametrized by the parameters λ, namely:

$$p_\lambda(x) = \Phi_\lambda \exp\left(\sum_{j\in[M]} \lambda_j f_j(x)\right), \tag{8}$$

[2] With little abuse of notation, we will use the same symbol for the probability distribution and its p.d.f., which we assume to exists w.r.t. a reference measure.

where $\Phi_\lambda := \int_{\mathcal{X}} \exp\left(\sum_{j\in[M]} \lambda_j f_j(x')\right) dx'$ is a normalization constant, which ensures that $p \in \Delta(\mathcal{X})$, and its log-transformation takes the name of log-partition function $A(\lambda) := \log \int_{\mathcal{X}} \exp(\sum_{j\in[M]} \lambda_j f_j(x)) dx$. The log-partition function defines the set of well-behaved distributions $\Omega = \{\lambda \in \mathbb{R}^M : A(\lambda) < +\infty\}$. At optimality, the parameters are defined as $\hat{\lambda}$ and correspond to the optimal Lagrangian multipliers of the dual of the empirical Max-Ent problem in Eq. (7). Now on, we will use $\hat{p}$ to identify $p_{\hat{\lambda}}$ for simplicity.

3 Distributional Policy Evaluation: A Max-Ent Approach

This section aims at answering the first research question Q1. The proposed approach turns distributional PE into a pure density estimation problem in a Max-Ent framework, called *Distributional Max-Ent Policy Evaluation*, as described in Algorithm 1. For this translation, the algorithm uses the distribution of returns η as p, N-trajectory samples $\mathcal{H}_N = \{\mathcal{H}\}_{n=0}^N$ as data, and a fixed set of features functions $\mathcal{F}$ belonging to a function class Φ. Note that to do this, we need to slightly change the notation concerning the D-RL framework: η will not be a $|\mathcal{S}|$-vector of distributions with support over $\mathbb{R}$, but rather a joint distribution over the whole support $\mathcal{X} = \mathcal{S} \times \mathbb{R}$. Turning PE into a Max-Ent problem has many upsides. First of all, the Max-Ent principle allows to deal with any kind of support $\mathcal{X}$, unifying continuous and discrete cases under the same framework; secondly, it does not require specifying a family of probability distributions to choose from; moreover, it implicitly manages the uncertainty by seeking a distribution as agnostic as possible, i.e., as close to the uniform distribution as possible. Finally, Max-Ent allows to include of structural constraints over the return distribution under many different flavors, both as in the standard value-function approximation methods [Van Roy, 2006] and as in more recent works based on statistical functionals acting over the return portion $\mathbb{R}$ of the support [Bellemare et al., 2023]. One of the possible limitations might be the requirement to have access to a batch of i.i.d. samples, but this is not necessarily restrictive: the result can be generalized for a single β-mixing sample path by exploiting blocking techniques [Yu, 1994, Nachum et al., 2019].

Algorithm 1 Distributional Max-Ent Policy Evaluation

Require: $(\mathcal{H}_N, \mathcal{F})$ ▷ N trajectory samples, set of features functions

$\hat{\eta} = \underset{\eta}{\operatorname{argmax}}\, H(\eta)$

s.t. $\mathbb{E}_{X\sim\eta}[f_j(X)] = \hat{\mu}_j(\mathcal{H}_N) \quad \forall j \in [M]$

$\eta \in \Delta(\mathcal{X})$

return $\hat{\eta}$

3.1 Generalization Error Bound

As previously said, the inner properties of Max-Ent allow for translating the results from density estimation methods to the distributional PE setting, and in particular, generalization-error bounds defined as KL-divergences.[3] Unfortunately, the generalization error bounds of traditional Max-Ent theory contain a conservative term that compares the solutions of the expectation and empirical Max-Ent problems, $\bar{\eta}, \hat{\eta}$ respectively, by taking the maximum between the 1-norm of the respective multipliers, namely $\max_{\lambda\in\{\bar{\lambda},\hat{\lambda}\}} ||\lambda||_1$. This quantity is bounded yet unknown, making the result unpractical. In the following, we extend the previous results with a more practical bound containing $||\hat{\lambda}||_1$ instead of the maximum, requiring some additional assumptions about the expressiveness of the feature functions. This result is of independent interest and allows us to directly use the bound from an algorithmic perspective.

Theorem 1 (Generalization Error Bound of D-Max-Ent PE)**.** *Assume that the set of features $\mathcal{F}$ belong to the function class Φ, which it is such that $\sup_{x\in\mathcal{X}, f\in\mathcal{F}} ||f(x)||_\infty = F < +\infty$ and that the minimum singular value $\sigma_{\min}$ of the empirical covariance matrix of the features $\hat{\mathrm{Cov}}(\mathcal{F})$ is strictly positive, namely $\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F})) > 0$. Then, given a sample batch $\{x_1, \dots, x_N\} \in \mathcal{X}^N$ of N i.i.d. points drawn from the true distribution η^π, for any $\delta \in (0, 1)$, it holds with probability at least $1 - \delta$*

[3] The KL-divergence between two distributions p, q is defined as $KL(p||q) = \mathbb{E}_p[\log(p/q)]$

that the solution to the sampled Max-Ent problem $\hat{\eta}$ satisfies the following:

$$KL(\eta^\pi||\hat{\eta}) \precsim -H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}) + B(\hat{\lambda}, \mathcal{F}, N, \delta) \tag{9}$$

$$\tilde{\mathcal{L}}(\hat{\eta}) = -\frac{1}{N}\sum_{i=0}^{N} \log \hat{\eta}(x_i) \tag{10}$$

$$B(\hat{\lambda}, \mathcal{F}, N, \delta) = 10||\hat{\lambda}||_1 \left(\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}} \right), \tag{11}$$

where $\precsim$ stands for the fact that the bound comprises additional terms that decrease at a higher rate in sample complexity and were therefore neglected. $H(\eta^\pi)$ and $\tilde{\mathcal{L}}(\hat{\eta})$, the empirical log-likelihood of the solution, form a bias term. The remaining term $B(\hat{\lambda}, \mathcal{F}, N, \delta)$ is a variance term depending on the multipliers characterizing the solution $\hat{\lambda}$, the number of samples, the confidence level δ, and the feature class complexity as the empirical Rademacher complexity of the class $\mathcal{R}_N(\Phi)$ [Mohri et al., 2018].

3.2 Proof Sketch

Here we report the main steps of the proof of Th. 1. The interested reader can find the complete proof in Appendix A. First, define the set containing the solutions to the expected and sampled Max-Ent problems with $\mathcal{S} := \{\bar{\eta}, \hat{\eta}\}$, the related set for the multipliers $\Omega_{\mathcal{S}} := \{\bar{\lambda}, \hat{\lambda}\}$, and a quantity that will be central now on $h(x_1, \cdots, x_N) := \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_i^N \log \eta(x_i)|$. Then, the building blocks of the error term $KL(\eta^\pi||\hat{\eta})$, namely $KL(\bar{\eta}||\hat{\eta})$ and $KL(\eta^\pi||\bar{\eta})$ are bounded by:

$$KL(\bar{\eta}||\hat{\eta}) \leq 2h(\cdot)$$
$$KL(\eta^\pi||\bar{\eta}) \leq -H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}) + 3h(\cdot).$$

It is possible to show that:

$$h(\cdot) \leq 2 \max_{\lambda \in \Omega_{\mathcal{S}}} ||\lambda||_1 \left(\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}} \right)$$
$$\max_{\lambda \in \Omega_{\mathcal{S}}} ||\lambda||_1 \leq ||\hat{\lambda}||_1 + \sqrt{\frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))} h(\cdot)}.$$

The first inequality is obtained with standard methods as in van der Vaart and Wellner [1996], Dudley [1999], Koltchinskii and Panchenko [2002], Wang et al. [2013]. The second one is obtained by exploiting the intrinsic properties of the Max-Ent solution and by noting that it is possible to link $h(\cdot)$ with the Bregman divergence of the log-partition function $D_A(\bar{\lambda}, \hat{\lambda})$. One can see that the use of the second inequality introduces an additional assumption about the expressiveness of the feature functions, requiring the minimum singular value of the sampled covariance matrix $\sigma_{\min}(\hat{Cov}(\mathcal{F}))$ to be strictly positive. As a final step, setting $x = \sqrt{h(x_1, \cdots, x_N)}$ and combining the two previous inequalities yields a quadratic inequality:

$$\begin{cases} x^2 - bx - c \leq 0 \\ b = 2\sqrt{\frac{6M}{\sigma_{\min}(\hat{Cov}(\mathcal{F}))}} \left[\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}} \right], \\ c = 2||\hat{\lambda}||_1 \left[\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}} \right] \end{cases}$$

which is well-defined and solves for

$$h(x_1, \cdots, x_N) \precsim ||\hat{\lambda}||_1 \left(\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}} \right),$$

by neglecting higher-order terms. The statement of the theorem is then just a matter of combining all these results.

4 Distributional Representation Learning with State Aggregation

This section addresses the second research question Q2, namely how to use the bound in Th. 1 from an algorithmic perspective to automatically refine the features used to represent the state space in a principled way while performing D-Max-Ent PE. In particular, the focus is on a specific instance of feature functions for return distributions, namely state aggregation. More specifically, the state aggregation feature functions $\mathcal{F} = \{f_j\}_{j\in[M]}$ split the state space into M disjoint subsets, one for each function, i.e., $\mathcal{S} = \cup_{j\in[M]} S_j$ and $S_j \cap S_{j'} = \emptyset,\ j, j' \in [M],\ j \neq j'$, and gives back the associated return $g \in \mathbb{R}$, namely:

$$\begin{aligned} &f_j : \mathcal{S} \times \mathbb{R} \to \mathbb{R} \\ &f_j(s, g) = g\mathbb{1}_{[s\in S_j]}. \end{aligned} \tag{12}$$

These features are bounded by the maximum return $G_{\max}$, while the empirical Rademacher complexity over N samples of returns $\{(s_i, g_i)\}_{i\in[N]}$ can be directly computed as in Clayton [2014]:

$$\mathcal{R}_N(\Phi) = G_{\max} \sum_{j\in[M]} \sqrt{\hat{P}(S_j)}, \tag{13}$$

where $\hat{P}(S_j) = N_j/N$ and $N_j = |\{(g_i, s_i) : s_i \in S_j, i \in [N]\}|$. The decomposition of the Rademacher term into single terms leads to rewriting $B(\hat{\lambda}, \mathcal{F}, N, \delta)$ as in the following lemma.

Lemma 1. *For Distributional Max-Ent Evaluation with a state-aggregation feature class, the variance term $B(\hat{\lambda}, \mathcal{F}, N, \delta)$ is given by;*

$$B(\hat{\lambda}, \mathcal{F}, N, \delta) = 10||\hat{\lambda}||_1 G_{max} \left(\sum_{j\in[M]} \sqrt{\hat{P}(S_j)} + \sqrt{\frac{\log 1/\delta}{2N}} \right). \tag{14}$$

4.1 Representation Refinement: Progressive Factorization

State aggregation features are of interest due to the possibility of progressively refining the representation by increasing the factorization level, that is, by splitting a subset S_j into further disjoint subsets. This refinement is called *progressive factorization* and is defined as follows.

Definition 1 (Progressive Factorization)**.** *For two sets of state aggregation feature functions, $\mathcal{F}, \mathcal{F}_j$, we say that $\mathcal{F}_j$ is a progressive factorization of $\mathcal{F}$, i.e., $\mathcal{F} \subset \mathcal{F}_j$, if $\mathcal{F} = \{f_1, \dots, f_{j-1}, f_{j+1}, \dots, f_M\} \cup \{f_j\}, \mathcal{F}_j = \{f_1, \dots, f_{j-1}, f_{j+1}, \dots, f_M\} \cup \{f_j^k\}_{k\in[K]}$ and the additional functions $\{f_j^k\}_{k\in[K]}$ are such that the corresponding subsets satisfy*

$$S_j = \bigcup_{k\in[K]} S_j^k, \quad S_j^k \cap S_j^{k'} = \emptyset,\ k, k' \in [K],\ k \neq k',$$

where only non-degenerate class factorizations will be considered, meaning that the new subsets S_j^k are non-empty.

It is relevant for our interests that, in the case of progressive factorizations $\mathcal{F} \subset \mathcal{F}'$, the respective Max-Ent solutions enjoy the following monotonicity property

Lemma 2 (Monotonicity)**.** *The multipliers of the Max-Ent solutions $\hat{\lambda}, \hat{\lambda}'$ using $\mathcal{F} \subset \mathcal{F}'$ are such that*

$$||\hat{\lambda}||_1 \leq ||\hat{\lambda}'||_1. \tag{15}$$

This result is fully derived in Appendix C, and it ensures a monotonically increasing of all terms contained in the variance term of Eq. (11) since the complexity term is monotonically increasing by definition. On the other hand, the bias represented by Eq. (10) is guaranteed to decrease monotonically at finer levels of factorizations.

4.2 D-Max-Ent Progressive Factorization Algorithm

In summary, D-Max-Ent PE shows a generalization error bound whose quantities are either known or estimated and change monotonically between progressive factorizations. On these results, we build an algorithm called *D-Max-Ent Progressive Factorization*, shown in Algorithm 2, which iteratively constructs a sequence of feature sets $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \dots$ with progressive factorization while performing PE. The behavior of the algorithm is similar to what is done in Structural Risk-Minimization (SRM) [Vapnik, 1991], and it involves optimizing for a trade-off: the bias term (i.e., empirical risk) decreases by taking into account more complex features classes, while the variance term (i.e., the confidence interval) increases. The whole algorithm is then based on the progressive search for the new set of feature functions which reduces a proxy of the generalization error bound of D-Max-Ent PE:

$$\mathcal{J}(\hat{\eta}) = \beta\mathcal{L}(\hat{\eta}) + B(\hat{\lambda}, \mathcal{F}, N, \delta), \qquad (16)$$

and the procedure will continue until there are no further improvements in the trade-off. Due to the nature of the proxy function, the role of $\beta > 0$ is to regulate the tendency to factorize. Higher values of β will increase the magnitude of the decreasing term, causing a boost in the tendency to factorize. On the other hand, lower values will further decrease the importance of this term, resulting in a lower tendency to factorization.

Algorithm 2 Distributional Max-Ent Progressive Factorization

Require: $(\mathcal{H}_N, \mathcal{F}_0, \delta, \beta, K)$ ▷ N-trajectory samples, initial feature set, confidence level, boosting factor, factorization factor

Done $\leftarrow$ False, $i^* \leftarrow 0$
while not Done **do**
  $\mathcal{F} \leftarrow \mathcal{F}_{i^*}, M \leftarrow |\mathcal{F}|$
  $\hat{\eta} \leftarrow$ D-Max-Ent PE$(\mathcal{H}_N, \mathcal{F})$
  $\mathcal{J}(\hat{\eta}) \leftarrow \beta\mathcal{L}(\hat{\eta}) + B(\hat{\lambda}, \mathcal{F}, N, \delta)$
  $\{\mathcal{F}_j\}_{j\in[M]} \leftarrow$ Progressive Factor$(\mathcal{F}, K)$
  for $j \in [M]$ **do**
    $\hat{\eta}_j \leftarrow$ D-Max-Ent PE$(\mathcal{H}_N, \mathcal{F}_j)$
    $\mathcal{J}(\hat{\eta}_j) \leftarrow \beta\mathcal{L}(\hat{\eta}_j) + B(\hat{\lambda}_j, \mathcal{F}_j, N, \delta)$
    if $\mathcal{J}(\hat{\eta}_j) < \mathcal{J}(\hat{\eta})$ **then**
      $i^* \leftarrow j$
    end if
  end for
  if $\mathcal{F}_{i^*} == \mathcal{F}$ **then**
    Done $\leftarrow$ True
  end if
end while
return $\hat{\eta}_{i^*}$

Finally, the *Progressive Factor* function takes as input a list of feature functions and a factoring factor K and returns a M-list of a progressively factored set of feature functions. More specifically, each element in $\{\mathcal{F}_j\}_{j\in[M]}$ progressively factors a different subset S_j into K disjoint subsets as in Definition 1. The new K subsets $\{\mathcal{S}_k^j\}_{k\in[K]}$ are constructed in the worst-case scenario: the complexity term in Eq. (13) is maximized with partitions of a set leading to a uniform distribution of samples in each new partitioned subset, and since it is not possible to know in advance which samples will be contained in which new subset, one way is then to proceed with a uniform factorization. We decided to maintain the most agnostic approach over the set of possible features, but prior knowledge could be used to narrow down the partitions to consider.

5 Illustrative Numerical Simulations

This section reports the results of some illustrative numerical simulations that make use of Algorithm 2. We have fully implemented the code base (out of the library used for Maximum Entropy Estimation) and we will make the software used in our experiments available as open-source.

Simulations Objectives The objectives of the simulations are twofold. First of all, to analyze the outcome of performing policy evaluations with aggregated states at different sample regimes, by comparing the output of the proposed algorithm with some relevant baseline distributions. Secondly, the aim is to study the role of the boosting parameter β and the sampling regime N, being the main hyper-parameters of Algorithm 2, in the tendency to factor the representation at utterly different sample regimes.

MDP Instance Design The effectiveness of the proposed approach is expected to be particularly evident in MDPs admitting a factored representation of the return distribution, namely the ones in which many states are nearly equivalent under the evaluation of a policy. This factorizability property is not uncommon in general since it is present in any environment with symmetries and Block-MDPs

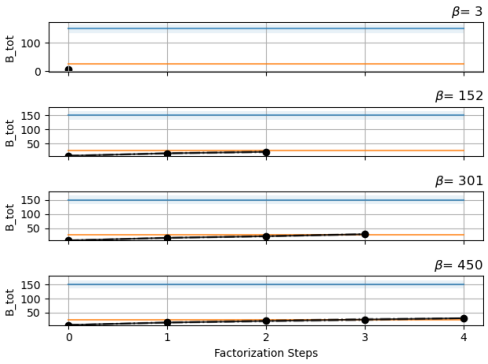

Figure 1: Bound Trend for different β ($N = 50$)

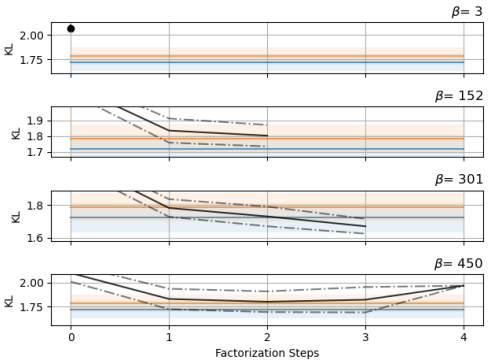

Figure 2: KL Trend for different β ($N = 50$)

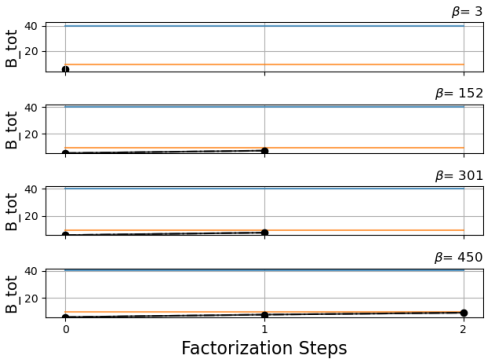

Figure 3: Bound Trend for different β ($N = 1000$)

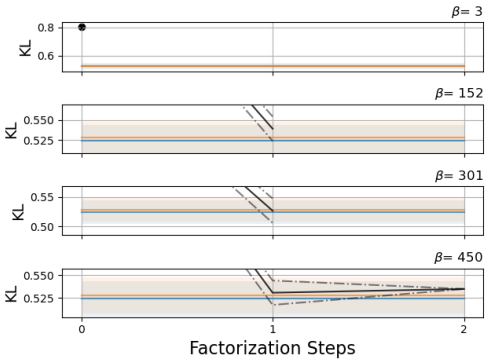

Figure 4: KL Trend for different β ($N = 1000$)

[Du et al., 2019] as well. The MDP instance is then designed to be a Block-MDP indeed since it allows for better evaluate the simulation objectives: one would expect that operating on MDPs admitting a factored representation would allow for lower values of β to be effective enough, while a higher level of boosting would force over-factorizations that are unnecessary, leading to no further improvement or even degradation of the results. The simulations are run on a rectangular GridWorld, with a height of 4 and length of 8, with traps on the whole second line and goals all over the top. A visualization of the setting can be seen in Appendix D. The policy is selected as a uniform distribution over the set of actions $\mathcal{A} = (\text{up, left, right})$.

Performance Indexes The proposed MDP instance presents many upsides in terms of the interpretability of the output as well. First of all, it allows us to directly compute the true underlying return distribution with Monte-Carlo estimation. Secondly, it permits to compare of the output distribution of the algorithm with the result of performing plain Distributional Max-Ent Policy Evaluation (Algorithm 1) with two baseline representations: an **oracle factorization** that aggregates together states known to be equivalent under the policy, and in particular all the upper and lower states respectively; a **full factorization** that employs $|\mathcal{S}|$-singletons of states as representations, i.e., the most fine-grained representation possible. The comparison is made via two relevant quantities, the KL divergence with respect to the true distribution (the *bounded quantity*), and the total bound $\mathrm{B}_{\text{tot}} = \tilde{\mathcal{L}}(\hat{\eta}) + B(\hat{\lambda}, \mathcal{F}, N, \delta)$ (the *bounding quantity*). Finally, the value of the partition splitting K is set to 2, to reduce the exponential search space of all possible uniform partitions, the discount factor γ is set to 0.98 and the confidence δ to 0.1, the results are averaged over 10 rounds with the respective standard deviation.

Results Discussion The results of the simulations are reported from Fig. 1 to Fig. 4, with the quantity related to the oracle parametrization being in **orange**, while the ones related to the full parametrization being in **blue**. Although these two distributions have almost the same KL divergence, they differ in the variance term in the bound, which is way higher for the case of full factorization. In all cases, it is

evident that the value of β pushes towards a higher level of factorizations, going from performing no factorization at all using low values ($\beta = 3, N \in \{50, 1000\}$), to performing up to 4 factorization steps even in this simple scenario ($\beta = 450, N = 50$). Furthermore, at higher sample regimes, the higher quality of the estimation counteracted the action of β, and increasing it generally induces fewer factorizations, as in Fig. 3, 4. Finally, it is apparent that minimizing for Eq. (16) successfully decreases the KL divergence. Nonetheless, its values stop decreasing significantly after the first factorization, which splits the state space over the two rows and further factorizations might lead to performance degradation as well.

6 Discussion

In this section, we briefly discuss the literature related to this work and provide some concluding remarks about the results and future research paths.

6.1 Related Works

Our work relates to multiple fields. We now highlight the most relevant connections, while an exhaustive overview is beyond the scope of this paper.

Distributional Reinforcement Learning D-RL has recently received much attention, both for the richness of flavors it admits [Rowland et al., 2019, 2021], and the surprising empirical effectiveness [Bellemare et al., 2020]. Our work tries to answer different research questions compared to traditional policy evaluation in D-RL and applies completely different techniques to derive the quantities of interest. In particular, it does not employ Bellman Operators or contraction arguments. In this way, the sequential nature of the problem is not exploited, but we show that density estimation techniques do offer interesting properties nonetheless. Additionally, the employed indexes differ from traditional D-RL results. For example, the most recent bound for Q-TD [Rowland et al., 2023], shows a sub-linear term but the bound is made over the maximum Wasserstein distance, which cannot be directly related to the KL-divergence without further assumptions. Finally, distributional considerations have been employed in the field of function approximations as well. However, to the best of our knowledge, no other D-RL works explicitly address representation learning.

Maximum Entropy and Feature Selection Max-Ent methods have a long and extensive literature in the density estimation field, which mostly focused on the general and algorithmic aspects of the method [Barron and Sheu, 1991, Dudík et al., 2004, Sutter et al., 2017]. Among the others, Cortes et al. [2015] proposed a Max-Ent regularized formulation for performing feature selection. Their method allocates different weights to different features to achieve an even better trade-off based on a combination. Due to this, their work differs from ours in the nature of the search space, which is not built progressively but is defined a priori. Additionally, their generalization bound is of the same nature as standard Max-Ent bounds and contains a $\sup_{\lambda \in \Omega} ||\lambda||_1$ term, which is bounded yet unknown. Finally, Mavridis et al. [2021] perform progressive state-aggregation through Max-Ent methods, but they try to optimize a different objective function based on state-dissimilarity.

6.2 Conclusions

In our work, we presented in a D-RL framework a new policy evaluation approach based on Maximum Entropy density estimation, called *Distributional Max-Ent* Policy Evaluation, which benefits from the learning guarantees of Max-Ent and the generality of the setting, being able to enforce even complex feature families. We extended previous results and derived a practical formulation of the generalization error bound, which contains only estimated and known quantities of the problem. We then instantiated a particular class of features, namely state aggregation, and we proposed an algorithm called *Distributional Max-Ent Progressive Factorization* to adaptively find a feature representation that optimizes for a proxy of the generalization error bound in a Structural Risk Minimization fashion. We then provided illustrative simulations showing the empirical behaviors of these approaches, while clarifying the links between some hyperparameters and the sample regime. Much of our analysis and theoretical guarantees straightforwardly extend to other feature classes, and an open question is to investigate other instances of features and settings that can benefit from the proposed framework. Another open question is whether there exist MDP instances that enjoy some relevant properties in the bias/variance trade-off along successive factorizations.

References

Andrew R. Barron and Chyong-Hwa Sheu. Approximation of Density Functions by Sequences of Exponential Families. *The Annals of Statistics*, 19(3), September 1991. ISSN 0090-5364. doi: 10.1214/aos/1176348252.

Marc G. Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C. Machado, Subhodeep Moitra, Sameera S. Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, December 2020. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-020-2939-8.

Marc G. Bellemare, Will Dabney, and Mark Rowland. *Distributional Reinforcement Learning*. MIT Press, 2023.

Daniel Brown, Scott Niekum, and Marek Petrik. Bayesian robust optimization for imitation learning. *Advances in Neural Information Processing Systems*, 33:2479–2491, 2020.

Yinlam Chow, M. Ghavamzadeh, Lucas Janson, and M. Pavone. Risk-constrained reinforcement learning with percentile risk criteria. *Journal Of Machine Learning Research*, 2015.

Scott Clayton. Lecture notes (topic 10) of eecs 598: Statistical learning theory, 2014.

Corinna Cortes, Vitaly Kuznetsov, M. Mohri, and Umar Syed. Structural Maxent Models. July 2015.

Shi Dong, Benjamin Van Roy, and Zhengyuan Zhou. Provably Efficient Reinforcement Learning with Aggregated States, February 2020. arXiv:1912.06366 [cs, math, stat].

Simon S Du, Akshay Krishnamurthy, Nan Jiang, Alekh Agarwal, Miroslav Dudík, and John Langford. Provably efficient RL with Rich Observations via Latent State Decoding. 2019.

R. M. Dudley. *Uniform Central Limit Theorems*. Cambridge University Press, 1 edition, July 1999. ISBN 978-0-521-46102-3 978-0-521-05221-4 978-0-511-66562-2. doi: 10.1017/CBO9780511665622.

Miroslav Dudík and Robert E. Schapire. Maximum Entropy Distribution Estimation with Generalized Regularization. In *Learning Theory*, volume 4005, pages 123–138. Springer Berlin Heidelberg, Berlin, Heidelberg, 2006. ISBN 978-3-540-35294-5 978-3-540-35296-9. doi: 10.1007/11776420_12. Series Title: Lecture Notes in Computer Science.

Miroslav Dudík, Steven J. Phillips, and Robert E. Schapire. Performance Guarantees for Regularized Maximum Entropy Density Estimation. In *Learning Theory*, volume 3120, pages 472–486. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004. ISBN 978-3-540-22282-8 978-3-540-27819-1. doi: 10.1007/978-3-540-27819-1_33. Series Title: Lecture Notes in Computer Science.

Ramtin Keramati, Christoph Dann, Alex Tamkin, and Emma Brunskill. Being Optimistic to Be Conservative: Quickly Learning a CVaR Policy. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(04):4436–4443, April 2020. doi: 10.1609/aaai.v34i04.5870.

V. Koltchinskii and D. Panchenko. Empirical Margin Distributions and Bounding the Generalization Error of Combined Classifiers. *The Annals of Statistics*, 30(1):1 – 50, 2002. doi: 10.1214/aos/1015362183.

Christos N. Mavridis, Nilesh Suriyarachchi, and John S. Baras. Maximum-Entropy Progressive State Aggregation for Reinforcement Learning. In *2021 60th IEEE Conference on Decision and Control (CDC)*, pages 5144–5149, Austin, TX, USA, December 2021. IEEE. ISBN 978-1-66543-659-5. doi: 10.1109/CDC45484.2021.9682927.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. Adaptive computation and machine learning. The MIT Press, Cambridge, Massachusetts, second edition edition, 2018. ISBN 978-0-262-03940-6.

Ofir Nachum, Yinlam Chow, Bo Dai, and Lihong Li. DualDICE: Behavior-Agnostic Estimation of Discounted Stationary Distribution Corrections, November 2019. URL `http://arxiv.org/abs/1906.04733`. arXiv:1906.04733 [cs, stat].

Martin L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. USA, 1st edition, 1994. ISBN 0471619779.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Rémi Munos, Marc G. Bellemare, and Will Dabney. Statistics and Samples in Distributional Reinforcement Learning. 2019. doi: 10.48550/ARXIV.1902.08102.

Mark Rowland, Shayegan Omidshafiei, Daniel Hennes, Will Dabney, Andrew Jaegle, Paul Muller, Julien Pérolat, and Karl Tuyls. Temporal Difference and Return Optimism in Cooperative Multi-Agent Reinforcement Learning. page 9, 2021.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, Georg Ostrovski, Anna Harutyunyan, Karl Tuyls, Marc G. Bellemare, and Will Dabney. An Analysis of Quantile Temporal-Difference Learning, January 2023. arXiv:2301.04462 [cs, stat].

Satinder Singh, Tommi Jaakkola, and Michael Jordan. Reinforcement Learning with Soft State Aggregation. In *Advances in Neural Information Processing Systems*, volume 7. MIT Press, 1994.

Tobias Sutter, David Sutter, Peyman Mohajerin Esfahani, and John Lygeros. Generalized maximum entropy estimation. 2017. doi: 10.48550/ARXIV.1708.07311.

Richard S Sutton and Andrew G Barto. *Reinforcement learning: An introduction*. MIT press, 2018.

Csaba Szepesvári. Algorithms for Reinforcement Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 4(1):1–103, January 2010. ISSN 1939-4608, 1939-4616.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer Series in Statistics. Springer New York, New York, NY, 1996. ISBN 978-1-4757-2547-6 978-1-4757-2545-2. doi: 10.1007/978-1-4757-2545-2.

Benjamin Van Roy. Performance Loss Bounds for Approximate Value Iteration with State Aggregation. *Mathematics of Operations Research*, 31(2):234–244, May 2006. ISSN 0364-765X, 1526-5471. doi: 10.1287/moor.1060.0188.

V. Vapnik. Principles of Risk Minimization for Learning Theory. In *Advances in Neural Information Processing Systems*, volume 4. Morgan-Kaufmann, 1991.

Martin J. Wainwright and Michael I. Jordan. Graphical Models, Exponential Families, and Variational Inference. *Foundations and Trends in Machine Learning*, 1:1–305, 2007. ISSN 1935-8237, 1935-8245. doi: 10.1561/2200000001.

Shaojun Wang, Russell Greiner, and Shaomin Wang. Consistency and Generalization Bounds for Maximum Entropy Density Estimation. *Entropy*, 15(12):5439–5463, December 2013. ISSN 1099-4300. doi: 10.3390/e15125439.

Bin Yu. Rates of Convergence for Empirical Processes of Stationary Mixing Sequences. *The Annals of Probability*, 22(1):94–116, 1994. ISSN 0091-1798. Publisher: Institute of Mathematical Statistics.

A Main Proof and Lemmas

In this section, we proceed in providing a proof of Theorem 1 of the main paper, together with some useful lemmas instrumental for proving it. Again, we define the set containing the solutions to the expected and sampled Max-Ent problems with $\mathcal{S} := \{\bar{\eta}, \hat{\eta}\}$, the related set for the multipliers $\Omega_{\mathcal{S}} := \{\bar{\lambda}, \hat{\lambda}\}$, which is a restriction of $\Omega = \{\lambda \in \mathbb{R}^M : A(\lambda) < +\infty\}$, and a quantity that will be central now on $h(x_1, \cdots, x_N) := \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_i^N \log \eta(x_i)|$.

Contribution Highlights The whole structure of the proof is built upon several intermediate results, of which some use standard techniques, and others are novel to this work. Here we report some comments to better clarify our contributions:

- Lemma 3 bounds the generalization-error with $h(\cdot)$, and it is based on the straighforward combination of Lemma 4 and Lemma 5.
- Lemma 4 introduces a slight modification to Wang et al. [2013] that is the use of the $\max_{\Omega_{\mathcal{S}}}$ over a finite set rather than $\sup_{\Omega}$ over the entire set of distributions. This will allow us to combine the result with the one of Lemma 5 and to deal with a simpler term, namely $h(x_1, \cdots, x_N)$ defined over the $\max$ instead of the $\sup$.
- Lemma 5 is a novel contribution, which was needed to obtain a practical form for the generalization error, compared to the intermediate result of Wang et al. [2013]. In this lemma as well $\max_{\Omega_{\mathcal{S}}}$ is employed, rather than $\sup_{\Omega}$.
- Lemma 6 uses standard techniques as can be found in van der Vaart and Wellner [1996], Dudley [1999], Koltchinskii and Panchenko [2002], but the analysis is again restricted to $\max_{\Omega_{\mathcal{S}}}$ thanks to the previous results.
- Lemma 7, Lemma 8 are novel results. They are needed to derive a practical generalization-error bound. Lemma 7 upper-bounds $||\bar{\lambda}||_1$ with $||\hat{\lambda}||_1$ by requiring additional constraints about the expressiveness of the feature functions. Lemma 8 uses this result to substitute $\max_{\lambda \in \Omega_{\mathcal{S}}} ||\lambda||_1$ with $||\hat{\lambda}||_1$.

As previously said, one of the main positives of this derivation is the ability to operate over $\max_{\eta \in \mathcal{S}}$ rather than $\sup_{\lambda \in \Omega}$. We will highlight the passages where this quantity is introduced with a $(\star)$, and provide further comments.

Initial step

First of all, we proceed in bounding the generalization error by bounding two sub-terms building it, that the following Lemma 3 will consist of a combination of two following lemmas, Lemma 4 and Lemma 5.

Lemma 3. *The generalization error between the true distribution and the Max-Ent solution of the sampled problem $\eta^\pi, \hat{\eta}$ (expressed as KL-divergence between the two distributions), given N i.i.d. samples, can be bounded with the following quantity:*

$$KL(\eta^\pi || \hat{\eta}) \leq -H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}) + 5 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$$

Proof. As said, the result directly follows by considering that for the problem under consideration $KL(\eta^\pi || \hat{\eta}) = KL(\bar{\eta} || \hat{\eta}) + KL(\eta^\pi || \bar{\eta})$, since the two solutions correspond to the exact and sampled estimation problems. To bound the term on the right it is sufficient to bound the two terms on the left. We know that according to Lemma 4,

$$KL(\bar{\eta} || \hat{\eta}) \leq 2 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$$

And according to Lemma 5

$$KL(\eta^\pi || \bar{\eta}) \leq -H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}) + 3 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$$

And the result directly follows. □

Lemma 4. *For the solutions of the exact and sampled Max-Ent problems, $\bar{\eta}$ and $\hat{\eta}$ respectively, it holds that*

$$KL(\bar{\eta}||\hat{\eta}) \leq 2 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)|$$

Proof.

$$\begin{aligned}
KL(\bar{\eta}||\hat{\eta}) &= KL(\eta^\pi||\hat{\eta}) - KL(\eta^\pi||\bar{\eta}) \\
&= (\mathbb{E}_{\eta^\pi}[\log \bar{\eta}] - \mathbb{E}_{\tilde{\eta}}[\log \bar{\eta}]) + (\mathbb{E}_{\tilde{\eta}}[\log \hat{\eta}] - \mathbb{E}_{\eta^\pi}[\log \hat{\eta}]) + (\mathbb{E}_{\tilde{\eta}}[\log \bar{\eta}] - \mathbb{E}_{\tilde{\eta}}[\log \hat{\eta}]) \\
&\leq 2 \max_{\eta \in \mathcal{S} = \{\bar{\eta}, \hat{\eta}\}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)| + \frac{1}{N} \sum_{j=0}^{N} \log \frac{\bar{\eta}(x_j)}{\hat{\eta}(x_j)} \quad (\star) \\
&\leq 2 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)|
\end{aligned}$$

where the term $\frac{1}{N} \sum_{j=0}^{N} \log \frac{\bar{\eta}(x_j)}{\hat{\eta}(x_j)}$ is negative and then is removed from the bounding scheme.

$(\star)$ Here, Wang et al. [2013] bounded conservatively the first two terms $(\mathbb{E}_{\eta^\pi}[\log \bar{\eta}] - \mathbb{E}_{\tilde{\eta}}[\log \bar{\eta}]) + (\mathbb{E}_{\tilde{\eta}}[\log \hat{\eta}] - \mathbb{E}_{\eta^\pi}[\log \hat{\eta}])$ with the $\sup_{\lambda \in \Omega}$, yet we notice that the only two quantities of interest between which we are asked to maximize over are in the $\max_{\eta \in \mathcal{S} = \{\bar{\eta}, \hat{\eta}\}}$. □

Lemma 5. *For the solutions of the Max-Ent problem in expectation $\bar{\eta}$ it is possible to bound the KL-divergence with respect to the true distribution η^π with the following quantity*

$$KL(\eta^\pi||\bar{\eta}) \leq -H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}) + 3 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)|$$

Proof.

$$\begin{aligned}
|\mathcal{L}_{\eta^\pi}(\bar{\eta}) - \tilde{\mathcal{L}}(\hat{\eta})| &= |\mathbb{E}_{\eta^\pi}[\log \bar{\eta}] - \frac{1}{N} \sum_{j=0}^{N} \log \hat{\eta}(x_j)| \\
&\leq |\mathbb{E}_{\eta^\pi}[\log \bar{\eta}] - \mathbb{E}_{\eta^\pi}[\log \hat{\eta}]| + |\mathbb{E}_{\eta^\pi}[\log \hat{\eta}] - \frac{1}{N} \sum_{j=0}^{N} \log \hat{\eta}(x_j)| \\
&\leq |KL(\eta^\pi||\bar{\eta}) - KL(\eta^\pi||\hat{\eta})| + \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)| \quad (\star) \\
&\leq 2 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)| + \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)| \\
&\leq 3 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)|
\end{aligned}$$

$(\star)$ Again, due to the conservative bound in Lemma 4, Wang et al. [2013] maintained the same quantity in this bound for later simplifications. We apply a tighter bound of $\max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N} \sum_{j=0}^{N} \log \eta(x_j)|$ to $|\mathbb{E}_{\eta^\pi}[\log \hat{\eta}] - \frac{1}{N} \sum_{j=0}^{N} \log \hat{\eta}(x_j)|$.

It follows that it is possible to write

$$|\mathcal{L}_{\eta^\pi}(\bar{\eta}) - \tilde{\mathcal{L}}(\hat{\eta})| = |KL(\eta^\pi||\bar{\eta}) + H(\eta^\pi) - \tilde{\mathcal{L}}(\hat{\eta})|$$

$$|KL(\eta^\pi||\bar{\eta}) - (-H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}))| \leq 3 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$$

$$||KL(\eta^\pi||\bar{\eta})| - |(-H(\eta^\pi) + \tilde{\mathcal{L}}(\hat{\eta}))|| \leq 3 \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$$

which proves the result. □

Intermediate Step

As suggested by the previous considerations, everything boils down to being able to bound the term $h(x_1, \cdots, x_N) \coloneqq \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$. To do this, we used standard techniques to derive the following intermediate step, where we can bound the quantity of interest which depends on the supremum between distributions $\max_{\eta \in \mathcal{S}} |\cdot|$ with a quantity depending on the supremum between their respective parameters $\lambda \in \Omega_{\mathcal{S}}$, namely $\sup_{\lambda \in \Omega_{\mathcal{S}}} ||\lambda||_1$.

Lemma 6. *The supremum difference between the expected log-likelihood and the sampled one, taken over the expected and sampled solutions in $\mathcal{S} = \{\bar{\lambda}, \hat{\lambda}\}$, is defined as $h(x_1, \cdots, x_N) \coloneqq \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$ and it can be bounded by*

$$\max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)| \leq 2 \sup_{\lambda \in \Omega_{\mathcal{S}}} ||\lambda||_1 \mathcal{R}_N(\Phi) + 2 \sup_{\lambda \in \Omega_{\mathcal{S}}} ||\lambda||_1 F \sqrt{\frac{\log 1/\delta}{2N}}$$

with $F = \sup_{f \in \mathcal{F}} ||f||_\infty$.

Proof. We define

$$h(x_1, \ldots, x_N) = \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|$$

$$= \sup_{\lambda \in \Omega_{\mathcal{S}}, f \in \mathcal{F}} |\mathbb{E}_{\eta^\pi} \langle \lambda, f(x) \rangle - \frac{1}{N}\sum_{j=0}^{N} \langle \lambda, f(x) \rangle|$$

Then by exploiting the definition of the function, we study the differences induced by changing one sample from x_k to x'_k

$$|h(x_1, \ldots, x_M) - h(x_1, \ldots, x'_k, \ldots, x_M)| =$$

$$= |\sup_{\lambda \in \Omega_{\mathcal{S}}, f \in \mathcal{F}} |\mathbb{E}_{\eta^\pi} \langle \lambda, f(x) \rangle - \frac{1}{N}\sum_{j=0}^{N} \langle \lambda, f(x) \rangle|$$

$$- \sup_{\lambda \in \Omega_{\mathcal{S}}, f \in \mathcal{F}} |\mathbb{E}_{\eta^\pi} \langle \lambda, f(x) \rangle - \frac{1}{N}\sum_{j \neq k}^{N} \langle \lambda, f(x) \rangle + \langle \lambda, f(x'_k) \rangle||$$

$$\leq \sup_{\lambda \in \Omega_{\mathcal{S}}, f \in \mathcal{F}} \frac{1}{N} |\langle \lambda, f(x_k) - f(x'_k) \rangle|$$

$$\leq \frac{2}{N} \sup_{\lambda \in \Omega_{\mathcal{S}}, f \in \mathcal{F}} ||\lambda||_1 ||f||_\infty = \frac{C}{N} \quad (C = 2 \sup_{\lambda \in \Omega_{\mathcal{S}}, f \in \mathcal{F}} ||\lambda||_1 ||f||_\infty)$$

Now, by Mc Diarmid's inequality, by studying the function concerning its sampled expectation $\mathbb{E}_{\tilde{\mathcal{X}}} h(\cdot)$ over the samples set $\tilde{\mathcal{X}} = \{x_1, \ldots, x_N\}$:

$$P(h(x_1, \ldots, x_N) - \mathbb{E}_{\tilde{\mathcal{X}}} h(x_1, \ldots, x'_k, \ldots, x_N) \geq \epsilon) \leq \exp(\frac{-2N\epsilon^2}{C^2})$$

$$P\Big(h(x_1, \ldots, x_N) - \mathbb{E}_{\tilde{\mathcal{X}}} h(x_1, \ldots, x'_k, \ldots, x_N) \geq C\sqrt{\frac{\log 1/\delta}{2N}}\Big) \leq \delta$$

It then follows that

$$\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)| \leq \mathbb{E}_{\tilde{\mathcal{X}}} \sup_{\lambda\in\Omega_{\mathcal{S}}, f\in\mathcal{F}} |\mathbb{E}_{\eta^\pi}\langle\lambda, f(x)\rangle - \frac{1}{N}\sum_{j=0}^{N}\langle\lambda, f(x)\rangle| + C\sqrt{\frac{\log 1/\delta}{2N}}$$

We now use symmetrization techniques by considering the Rademacher sequence $\{\omega_j\}$ and by using the standard result that given a class of measurable functions $\mathcal{G}$ if

$$Z(\tilde{\mathcal{X}}) = \sup_{g\in\mathcal{G}} |\mathbb{E}g(x) - \frac{1}{N}\sum_{j=0}^{N} g(x_j)| \quad \text{and } R(\tilde{\mathcal{X}}, \omega) = \sup_{g\in\mathcal{G}} |\frac{1}{N}\sum_{j=0}^{N}\omega_j g(x_j)|$$

Then:

$$\mathbb{E}_{\tilde{\mathcal{X}}} Z(\tilde{\mathcal{X}}) \leq 2\mathbb{E}_{\tilde{\mathcal{X}},\omega} R(\tilde{\mathcal{X}})$$

From this, it follows that the whole expression reduces to

$$\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)| \leq 2\mathbb{E}_{\tilde{\mathcal{X}},\omega} \sup_{\lambda\in\Omega_{\mathcal{S}}, f\in\mathcal{F}} |\frac{1}{N}\sum_{j=0}^{N}\omega_j\langle\lambda, f(x_j)\rangle| + C\sqrt{\frac{\log 1/\delta}{2N}}$$

We extract the supremum over $\lambda \in \Omega_{\mathcal{S}}$ to obtain the (absolute) Rademacher averages of the functions in $\mathcal{F}$

$$\begin{aligned}\mathbb{E}_\omega \sup_{\lambda\in\Omega_{\mathcal{S}}, f\in\mathcal{F}} |\frac{1}{N}\sum_{j=0}^{N}\omega_j\langle\lambda, f(x_j)\rangle| &\leq \sup_{\lambda\in\Omega_{\mathcal{S}}} ||\lambda||_1 \mathbb{E}_\omega \sup_{f\in\mathcal{F}} |\frac{1}{N}\sum_{j=0}^{N}\omega_j f(x_j)| \\ &\leq \sup_{\lambda\in\Omega_{\mathcal{S}}} ||\lambda||_1 \mathcal{R}_N(\Phi)\end{aligned}$$

It follows that the final formulation for the term we are studying is the following

$$\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)| \leq 2\sup_{\lambda\in\Omega_{\mathcal{S}}} ||\lambda||_1 \mathcal{R}_N(\Phi) + C\sqrt{\frac{\log 1/\delta}{2N}}$$

$$C = 2\sup_{\lambda\in\Omega_{\mathcal{S}}, f\in\mathcal{F}} ||\lambda||_1 ||f||_\infty$$

□

Final Step

The bound offered by Lemma 6 would be unpractical since it relates a quantity central to our analysis to something which is not known in advance. Due to this, we make a further effort with the following Lemma, by substituting the term $\sup_{\lambda\in\Omega_{\mathcal{S}}} ||\lambda||_1$ with $||\hat{\lambda}||_1$. To do this, an additional assumption over the feature functions will be needed though. First of all, we bound the two terms in $\Omega_{\mathcal{S}}$ with

Lemma 7. *The solutions of the expected and sampled Max-Ent problem are related to the bound:*

$$||\bar{\lambda}||_1 \leq ||\hat{\lambda}||_1 + \sqrt{\frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))} \max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)|}$$

Proof. We take into account the following relationships which are valid for the solutions of the MaxEnt problem under structural constraints, i.e. $\mathbb{E}_{\bar{\eta}}[f] = \mathbb{E}_{\eta^\pi}[f]$ and $\mathbb{E}_{\hat{\eta}}[f] = \mathbb{E}_{\tilde{\eta}}[f]$

$$\begin{aligned}H(\eta) &= \log\sum_y \exp(\langle\lambda, f(y)\rangle) - \langle\lambda, \mathbb{E}_\eta[f]\rangle \\ &= A(\lambda) - \langle\lambda, \mathbb{E}_\eta[f]\rangle = A(\lambda) - \langle\lambda, \nabla A(\lambda)\rangle\end{aligned}$$

From which it follows that it is possible to recover the Bregman divergence under the log-partition function $D_A(\lambda_1, \lambda_2)$

$$\begin{aligned}
H(\bar{\eta}) - H(\hat{\eta}) &= A(\bar{\lambda}) - A(\hat{\lambda}) - \langle \bar{\lambda}, \nabla A(\bar{\lambda}) \rangle + \langle \hat{\lambda}, \nabla A(\hat{\eta}) \rangle \\
&= A(\bar{\lambda}) - A(\hat{\lambda}) - \langle \bar{\lambda}, \nabla A(\bar{\lambda}) \rangle + \langle \hat{\lambda}, \nabla A(\hat{\lambda}) \rangle + \langle \bar{\lambda}, \nabla A(\hat{\lambda}) \rangle - \langle \bar{\lambda}, \nabla A(\hat{\eta}) \rangle \\
&= A(\bar{\lambda}) - A(\hat{\lambda}) - \langle \bar{\lambda} - \hat{\lambda}, \nabla A(\hat{\lambda}) \rangle + \langle \hat{\lambda}, \nabla A(\hat{\lambda}) - \nabla A(\bar{\lambda}) \rangle \\
&= D_A(\bar{\lambda}, \hat{\lambda}) + \langle \bar{\lambda}, \nabla A(\hat{\lambda}) - \nabla A(\bar{\lambda}) \rangle
\end{aligned}$$

Now using the Taylor expansion of the divergence and the fact that $\nabla^2 A(\hat{\lambda}) = \hat{\mathrm{Cov}}(\mathcal{F})$

$$\begin{aligned}
H(\bar{\eta}) - H(\hat{\eta}) + \langle \bar{\lambda}, \nabla A(\bar{\lambda}) - \nabla A(\hat{\lambda}) \rangle &= D_A(\bar{\lambda}, \hat{\lambda}) \\
&\geq \frac{1}{2}(\bar{\lambda} - \hat{\lambda})^\intercal \nabla^2 A(\hat{\lambda})(\bar{\lambda} - \hat{\lambda}) = \frac{1}{2}||\bar{\lambda} - \hat{\lambda}||^2_{\nabla^2 A(\hat{\lambda})} \\
&\geq \sigma_{\min}(\nabla^2 A(\hat{\lambda}))||\bar{\lambda} - \hat{\lambda}||_2^2 \\
&\geq \frac{\sigma_{\min}(\nabla^2 A(\hat{\lambda}))}{M}||\bar{\lambda} - \hat{\lambda}||_1^2 \\
&\geq \frac{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))}{M}||\bar{\lambda} - \hat{\lambda}||_1^2
\end{aligned}$$

where M corresponds to the number of the features. Finally, by exploiting the zero duality gap and the results of Lemma 10

$$\begin{aligned}
||\bar{\lambda} - \hat{\lambda}||_1^2 &\leq \frac{M}{\sigma_{\min}(\mathrm{Cov}_{\hat{\lambda}}(f))}(H(\bar{\eta}) - H(\hat{\eta}) + \langle \bar{\lambda}, \nabla A(\bar{\lambda}) - \nabla A(\hat{\lambda}) \rangle) \\
&= \frac{M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))}(\mathcal{L}_0(\bar{\lambda}) - \tilde{\mathcal{L}}(\hat{\lambda}) + \langle \bar{\lambda}, \nabla A(\bar{\lambda}) - \nabla A(\hat{\lambda}) \rangle) \\
&\leq \frac{M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))}(|\mathcal{L}_0(\bar{\lambda}) - \tilde{\mathcal{L}}(\hat{\lambda})| + |\langle \bar{\lambda}, \nabla A(\bar{\lambda}) - \nabla A(\hat{\lambda}) \rangle|) \\
&\leq \frac{2M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))}|\mathcal{L}_0(\bar{\lambda}) - \tilde{\mathcal{L}}(\hat{\lambda})| \\
&\leq \frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))} \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|
\end{aligned}$$

It is then possible to write

$$||\bar{\lambda} - \hat{\lambda}||_1 \leq \sqrt{\frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))} \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|}$$

$$|||\bar{\lambda}||_1 - ||\hat{\lambda}||_1| \leq ||\bar{\lambda} - \hat{\lambda}||_1 \leq \sqrt{\frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))} \max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)|}$$

which concludes the proof. □

Now, it is possible to combine all the previous results in

Lemma 8. *Assume that the minimum singular value of the sampled covariance matrix is strictly positive, that is $\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F})) > 0$, then the supremum term of Lemma 6 can be bounded with*

$$\max_{\eta \in \mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log \eta] - \frac{1}{N}\sum_{j=0}^{N} \log \eta(x_j)| \precsim 2||\hat{\lambda}||_1 \mathcal{R}_N(\Phi) + 2||\hat{\lambda}||_1 F \sqrt{\frac{\log 1/\delta}{2N}}$$

Proof. Taking all together the terms obtained so far from Lemmas [6, 7], setting $C = 2\sup_{\lambda\in\{\bar{\lambda},\hat{\lambda}\},f\in\mathcal{F}} ||\lambda||_1||f||_\infty$ we have

$$\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)| \leq 2\sup_{\lambda\in\{\bar{\lambda},\hat{\lambda}\}} ||\lambda||_1\mathcal{R}_N(\Phi) + C\sqrt{\frac{\log 1/\delta}{2N}}$$

$$\sup_{\lambda\in\{\bar{\lambda},\hat{\lambda}\}} ||\lambda||_1 \leq ||\hat{\lambda}||_1 + \sqrt{\frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))}\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)|}$$

It follows the quadratic form in $x = \sqrt{\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)|}$

$$x^2 - bx - c \leq 0$$

$$b = 2\sqrt{\frac{6M}{\sigma_{\min}(\hat{\mathrm{Cov}}(\mathcal{F}))}}\left[\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}}\right] \geq 0$$

$$c = 2||\hat{\lambda}||_1\left[\mathcal{R}_N(\Phi) + F\sqrt{\frac{\log 1/\delta}{2N}}\right] \geq 0$$

The discriminant is well defined $\Delta = b^2 + 4c \geq 0$ and the solution is given by

$$\begin{aligned}\max_{\eta\in\mathcal{S}} |\mathbb{E}_{\eta^\pi}[\log\eta] - \frac{1}{N}\sum_{j=0}^{N}\log\eta(x_j)| &\leq \Big(\frac{b+\sqrt{b^2+4c}}{2}\Big)^2 \\ &\leq \frac{b^2}{2} + c + b\sqrt{b^2+4c} \\ &\lesssim 2||\hat{\lambda}||_1\mathcal{R}_N(\Phi) + 2||\hat{\lambda}||_1 F\sqrt{\frac{\log 1/\delta}{2N}}\end{aligned}$$

The final step was done because all additional terms out of c itself are of higher order. □

B Further instrumental Lemmas

In this section, we present some additional standard lemmas which summarize some important properties of the Max-Ent solutions and distributions in the exponential family that was used in the employed section.

Lemma 9. *For any distribution η in the exponential family, it holds that for the log-likelihood with respect to a distribution η^π it holds that*

$$\mathcal{L}_{\eta^\pi}(\lambda) = A(\lambda) - \langle\lambda, \mathbb{E}_{\eta^\pi}[f]\rangle$$

Proof.

$$\mathcal{L}_{\eta^\pi}(\lambda) = -\mathbb{E}_{\eta^\pi}[\log\eta] = -\mathbb{E}_{\eta^\pi}[\langle\lambda, f\rangle - \log\Phi_\lambda] = -\langle\lambda, \mathbb{E}_{\eta^\pi}[f]\rangle + A(\lambda)$$

□

Lemma 10. *For any distribution η in the exponential family, it holds that*

$$|\mathcal{L}_{\eta^\pi}(\lambda) - \tilde{\mathcal{L}}(\lambda)| = |\langle\lambda, \mathbb{E}_{\eta^\pi}[f] - \tilde{\mathbb{E}}[f]\rangle|$$

where $\mathcal{L}_{\eta^\pi}(\lambda)$ is the negative log-likelihood of η with respect to η^π.

Proof.

$$\begin{aligned}|\mathcal{L}_{\eta^\pi}(\lambda) - \tilde{\mathcal{L}}(\lambda)| &= |-\langle\lambda, \mathbb{E}_{\eta^\pi}[f]\rangle + A(\lambda) + \langle\lambda, \mathbb{E}_{\tilde{\eta}}[f]\rangle - A(\lambda)| \\ &= |\langle\lambda, -\mathbb{E}_{\eta^\pi}[f] + \mathbb{E}_{\tilde{\eta}}[f]\rangle| \\ &= |\langle\lambda, \mathbb{E}_{\eta^\pi}[f] - \mathbb{E}_{\tilde{\eta}}[f]\rangle|\end{aligned}$$

□

We will now derive some properties between the sampled log-likelihood and the log-likelihood with respect to the true distribution η^π, called $\mathcal{L}_0$ for simplicity

Lemma 11. *For the solutions of the exact and sampled Max-Ent problems, $\bar{\eta}$ and $\hat{\eta}$ respectively, it holds that*

$$\begin{aligned}|\mathcal{L}_0(\bar{\lambda}) - \tilde{\mathcal{L}}(\hat{\lambda})| &\leq |\langle \hat{\lambda}, \mathbb{E}_{\eta^\pi}[f] - \mathbb{E}_{\tilde{\eta}}[f]\rangle| \\ &\leq |\langle \hat{\lambda}, \nabla A(\bar{\lambda}) - \nabla A(\hat{\lambda})\rangle|\end{aligned}$$

$$\begin{aligned}|\mathcal{L}_0(\bar{\lambda}) - \tilde{\mathcal{L}}(\hat{\eta})| &\geq |\langle \bar{\lambda}, \mathbb{E}_{\eta^\pi}[f] - \mathbb{E}_{\tilde{\eta}}[f]\rangle| \\ &\geq |\langle \bar{\lambda}, \nabla A(\bar{\lambda}) - \nabla A(\hat{\lambda})\rangle|\end{aligned}$$

Proof. The proof follows directly from the fact that $\bar{\lambda}$ is optimal with respect to $\hat{\eta}$ in the exact problem $\mathcal{L}_0(\bar{\lambda}) \leq \mathcal{L}_0(\hat{\lambda})$ and viceversa $\tilde{\mathcal{L}}(\bar{\lambda}) \geq \tilde{\mathcal{L}}(\hat{\eta})$. □

C Monotonicity Lemma

In this section, we provide the proof of Lemma 2.

Proof. Taking into account two features with increased factorization $\mathcal{F} \subset \mathcal{F}'$ we consider the particular set of factorized features $\bar{\alpha}, \{\bar{\alpha}_k\}$, since the rest of the features are the same. It follows that

$$\begin{aligned}|\hat{\mu}_{\bar{\alpha}}| &= \sum_k |\hat{\mu}_{\bar{\alpha}_k}| \\ \frac{|\hat{\mu}_{\bar{\alpha}}|}{|S_{\bar{\alpha}}|} &= \sum_k \frac{|\hat{\mu}_{\bar{\alpha}_k}|}{|S_{\bar{\alpha}}|} \quad (\forall \bar{\alpha}_k : |S_{\bar{\alpha}_k}| < |S_{\bar{\alpha}}|) \\ \frac{|\hat{\mu}_{\bar{\alpha}}|}{|S_{\bar{\alpha}}|} &\leq \sum_k \frac{|\hat{\mu}_{\bar{\alpha}_k}|}{|S_{\bar{\alpha}_k}|}\end{aligned}$$

Now, due to the relationship of Lemma 12 we know that $\hat{\lambda}_\alpha = f(\frac{|\hat{\mu}_{\bar{\alpha}}|}{|S_{\bar{\alpha}}|})$ with $f(\cdot)$ being an unknown but subadditive for positive values of λ. Moreover, the functions are the same for all the terms, so that

$$\begin{aligned}\frac{|\hat{\mu}_{\bar{\alpha}}|}{|S_{\bar{\alpha}}|} &\leq \sum_k \frac{|\hat{\mu}_{\bar{\alpha}_k}|}{|S_{\bar{\alpha}_k}|} \\ f(\frac{|\hat{\mu}_{\bar{\alpha}}|}{|S_{\bar{\alpha}}|}) &\leq f(\sum_k \frac{|\hat{\mu}_{\bar{\alpha}_k}|}{|S_{\bar{\alpha}_k}|}) \leq \sum_k f(\frac{|\hat{\mu}_{\bar{\alpha}_k}|}{|S_{\bar{\alpha}_k}|}) \\ |\lambda_{\bar{\alpha}}| &\leq \sum_k |\lambda_{\bar{\alpha}_k}|\end{aligned}$$

Since the rest of the terms are the same, this concludes the proof. □

Lemma 12. *There exists a monotonic and anti-symmetric function $f(\cdot)$ such that it is possible to univocally define $\hat{\lambda}_\alpha = f(\hat{\mu}_\alpha, |S_\alpha|, G_{max})$*

Proof. We start by considering the Lagrangian formulation of the Max-Ent problem,

$$\mathcal{L}(\eta, \lambda) = H(\eta) + \sum_{\alpha \in \mathcal{I}_\mathcal{F}} \lambda_\alpha(\mathbb{E}_\eta[f_\alpha] - \hat{\mu}_\alpha) + \mu(\mathbb{E}[\eta] - 1) \tag{17}$$

By taking the gradient of the Lagrangian with respect to the distribution it follows that each x-term of the support gives

$$(\nabla_\eta \mathcal{L})(x) = -1 - \log \eta(x) + \lambda_\alpha f_\alpha + \mu$$

From which it follows that with $\lambda_0 = \mu - 1$ the equation for the α-constraint is

$$\eta_\alpha(x) = e^{\lambda_0} e^{\lambda_\alpha f_\alpha(x)}$$

We now compute insert this value inside the constraint equation under the feature class $f_\alpha = g\mathbb{1}_{s \in \mathcal{S}_\alpha}$

$$\int_{\mathcal{R}} \int_{\mathcal{X}} g\eta(x)_\alpha = \hat{\mu}_\alpha$$

$$|S_\alpha| \int_{G_{\text{min}}}^{G_{\text{max}}} g e^{\lambda_0} e^{\lambda_\alpha r} dg = \hat{\mu}_\alpha$$

which leads to the implicit formulation for λ_α by solving the integral by setting $G = G_{\text{max}}$

$$e^{\lambda_0} \frac{2\lambda_\alpha \cosh(G\lambda_\alpha) - 2\sinh(G\lambda_\alpha))}{\lambda_\alpha^2} = \frac{\hat{\mu}_\alpha}{|S_\alpha|}$$

Now, it can be proven by considering the normalization constraint that $e^{\lambda_0} = 1/Z(\lambda)$ with $Z(\lambda)$ a constant depending on λ_α, in particular:

$$\begin{aligned} Z(\lambda) &= \int_{\mathcal{X}} e^{\sum_\alpha \lambda_\alpha f_\alpha} dx \\ &= \sum_\alpha |S_\alpha| \int_{\mathcal{R}} e^{\lambda_\alpha f_\alpha} dr \\ &= \sum_\alpha |S_\alpha| \frac{\sinh \lambda_\alpha G}{\lambda_\alpha} \\ &= |S_\alpha| \frac{\sinh \lambda_\alpha G}{\lambda_\alpha} + C \end{aligned}$$

The whole equation then becomes

$$\frac{2\lambda_\alpha \cosh(G\lambda_\alpha) - 2\sinh(G\lambda_\alpha))}{\lambda_\alpha^2} = \hat{\mu}_\alpha (\frac{\sinh \lambda_\alpha G}{\lambda_\alpha} + C)$$

This equation provides an implicit definition for λ_α and it can be shown to be convex for positive values of λ. The function for lambda is the inverse of this whole term, which is then concave and has a zero in the origin, thus it is sub-additive. □

D Further Discussion about the Environment Setup

The GridWorld instance was selected to be able to satisfy the Experiment Objectives. More specifically, for the following reasons:

- be able to test the representation learning capacity of the algorithm in a setting with an apparent factorization of the state space
- because of the possibility of explicitly comparing the outputs with the true distribution.

As previously said, the simulations were run on a rectangular GridWorld, with a height of 4 and length of 8, with traps on the whole second line and goals all over the top. A visualization of the Gridworld can be seen in Fig.5.

E Reproducibility

The code was run over a 4 core Intel(R) Core(TM) i5-4570 CPU @ 3.20GHz. The package use to perform Maximum Entropy density estimation can be found at Max-Ent. The repository can be found at the anonymized link Code.

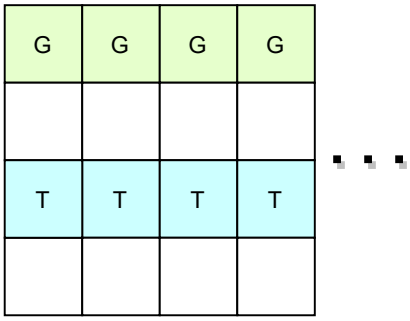

Figure 5: GridWorld template