An analysis of distributional reinforcement learning with Gaussian mixtures

Mathis Antonetti *mathis.antonetti@inria.fr*
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Henrique Donâncio *henrique.donancio@inria.fr*
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Florence Forbes *florence.forbes@inria.fr*
Univ. Grenoble Alpes, Inria, CNRS, Grenoble INP, LJK, 38000 Grenoble, France.

Reviewed on OpenReview: *https://openreview.net/forum?id=b4VgI1RTv8*

Abstract

Distributional Reinforcement Learning (DRL) seeks to optimize risk-sensitive objectives by modeling the full return distribution rather than only its expectation. A key challenge is to choose a return distribution representation that allows (i) efficient estimation of risk measures, (ii) tractable optimization, and (iii) sufficient expressiveness. Gaussian mixtures (GM) provide a flexible and powerful representation for this purpose, yet they remain underexplored in DRL, with most existing methods relying on the L_2 norm as a tractable metric between GM. In this work, we conduct a theoretical and empirical study of alternative metrics for GM-based DRL. We show that the L_2 norm is not suitable and introduce two principled alternatives: a mixture-specific optimal transport distance (MW) and a maximum mean discrepancy (MMD) distance. For the MW metric, we establish convergence guarantees for a dynamic programming algorithm related to temporal-difference (TD) learning. Leveraging multivariate GM representations, we also highlight the potential of MW in multi-objective RL. Experimental results on selected Atari Learning Environment tasks illustrate the practical benefits of the proposed metrics, showing promising performance.

1 Introduction

Deep Reinforcement Learning (RL) has shown outstanding results in robotics (Li et al., 2018), and control (Yang et al., 2018; Kaufmann et al., 2023), by estimating and optimizing the expectation of the total reward (or return) given to an agent. Distributional RL (DRL) (Bellemare et al., 2023) generalizes the approach by estimating the whole distribution of the return, leading to better results than the non-distributional approach (Dabney et al., 2018a). DRL has many other benefits. It fits better than the traditional approach in a stochastic setting (Martin et al., 2020) and it allows for new exploration strategies taking into account aleatoric uncertainty (Mavrin et al., 2019). DRL can also be used to design risk-aware RL agents using distortion risk measures that could lead to better results depending on the environment. Indeed, risk-averse policies lead to longer play times and better returns, in life-dependent games (Dabney et al., 2018a).

To derive practical DRL algorithms, two ingredients are required. First, tractable and informative representations of returns as probability distributions, which are infinite dimensional objects. Second, an efficient quantitative way to compare them through tractable metrics, to define a relevant and useful loss, based on a so-called distributional Bellman operator and typically similar to a temporal difference (TD) learning loss in RL. The metric choice is also crucial since each metric offers different theoretical convergence guarantees, *e.g.* to characterize the contraction property of the distributional Bellman operator and its projection.

Choosing an appropriate representation of the return distribution remains challenging as it has to meet somewhat opposite requirements. The representation has to be rich enough to capture the return distribution complexity and simple enough to allow tractable implementations at reasonable computing costs. Current state-of-the-art DRL methods need large computational resources and/or are mainly based on convenient choices with no performance guarantee in terms of their ability to represent actual return distributions. For instance, some methods use discrete distribution representations (C51 in Bellemare et al. (2017a), QR-DQN in Dabney et al. (2018b)) leading to low approximation rates and consequently large numbers of parameters, which increase the computational cost. Another theoretical downside is that those algorithms learn statistics of the distribution and not the distribution itself. It follows that they need to satisfy the approximate Bellman-closedness property as most, such as C51, do not satisfy the exact Bellman-closedness property (Rowland et al., 2019). Another set of methods use continuous representations (IQN by Dabney et al. (2018a), FQF by Yang et al. (2019)) that do not lead to a proper distribution in practice as they approximate the quantile function with non-monotonic surrogates, which questions their use with a risk-aware agent (Théate et al., 2023). More generally, most approaches use nonparametric empirical measure representations as opposed to parametric ones corresponding to a family of parametric distributions.

In this paper, we investigate parametric representations and more specifically Gaussian mixtures (GM). Introduced in the DRL framework by Choi et al. (2019), Gaussian mixtures present interesting features. They often lead to closed-form formulas, more likely to yield tractable loss and risk measure estimations. They are proper distributions, available in any dimension and with good approximation power, see *e.g.* Nguyen et al. (2023) for a recent reference. When it comes to computing metrics, there exists a number of closed-form expressions or efficient estimators specifically tuned to handle comparison between GM. Surprisingly, the large variety of tractable divergences between Gaussian mixtures has not been fully exploited yet in a DRL context. In this work, we consider three such metrics, the Jensen-Tsallis (JT), the maximum mean discrepancy (MMD) and a mixture-specific optimal transport distance, introduced by Delon & Desolneux (2020) and named the Mixture-Wasserstein (MW) distance. We compare them in terms of their theoretical and practical performance. More specifically, our main contributions can be summarized as follows:

- We prove that, although the JT metric has been used in DRL with GM (Choi et al., 2019), it does not guarantee that the distributional Bellman operator is a contraction mapping, leading to poor results for certain environments and justifying the search for alternatives. We propose an extended JT formulation with more flexibility.
- We introduce MW in DRL as a new possible metric, which is more tractable than the classical Wasserstein distance when comparing mixtures. We prove the contraction of the projected distributional Bellman operator with MW and the convergence of a dynamic programming algorithm related to TD learning. We then provide a generalization to multi-objective RL.
- We study various MMD kernels in the GM setting and give new insights on their performance in DRL. We observe that the so-called unrectified kernel performs as well as other common choices such as the mixture of Gaussian kernels. This is consistent with the good theoretical properties of the former but contrasts with previous results by Nguyen-Tang et al. (2020).

Overall, we address a critical problem in DRL, which is the search for effective metrics that maintain theoretical tractability and improve performance. We introduce a new bridge between classical mixture models and modern reinforcement learning, by revisiting GM through the lens of new probability metrics. We provide extensive theoretical justification (Table 1, Theorems 1 to 10) for these new metrics and a practical implementation framework. These metrics have not been studied in prior work in GM-based DRL (Choi et al., 2019; Nam et al., 2021; Zhang et al., 2024). To our knowledge, we present the most broad study of probability metrics under which it is feasible to learn GM representations of return distributions. In terms of impact, our work shows that although the most natural, the JT metric should not be used. Also, we propose the first alternative to the categorical approach, which comes with proved convergence and straightforward extensions to multi-objective DRL.

2 Distributional RL

In the standard RL setting, a Markov decision process $(\mathcal{X}, \mathcal{A}, R, P, \gamma)$ models the interaction between an agent and an environment. $\mathcal{X}$ and $\mathcal{A}$ denote the state and action spaces, $R(x,a)$ is a reward random variable depending on a given state $x \in \mathcal{X}$ and action $a \in \mathcal{A}$ with distribution $\rho(x,a)(\cdot)$, $P(\cdot|x,a)$ is a transition kernel to a new state from state x after taking action a, and $\gamma \in (0,1)$ is a discount factor. In RL, we search for a policy $\pi(\cdot|x)$ that maps a state x to a distribution over actions in $\mathcal{A}$. To emphasize this, we use upper case to denote random variables and lower case for their realizations. For an agent taking actions given by a policy π, the return is the random variable denoted by $Z^\pi(x,a) = \sum_{t=0}^{\infty} \gamma^t R(X_t, A_t)$ where $X_0 = x$, $A_0 = a$, $X_t \sim P(\cdot \mid x_{t-1}, a_{t-1}), A_t \sim \pi(\cdot \mid x_t)$. The Q-value function is defined by the expected return $Q^\pi(x,a) = \mathbb{E}[Z^\pi(x,a)]$ and can be characterized by the Bellman equation, $Q^\pi(x,a) = \mathbb{E}_{\rho(x,a)}[R(x,a)] + \gamma \mathbb{E}_{P(\cdot|x,a),\pi}[Q^\pi(X', A')]$, following from the random transition from (x,a) to realizations of (X', A') given by $X' \sim P(\cdot \mid x,a), A' \sim \pi(\cdot \mid x)$. The objective is then to find the optimal policy π^* which maximizes the expected return, with $Q^{\pi^*}(x,a) \geq Q^\pi(x,a)$ for all (x,a) and π. Q^{π^*} satisfies the optimal Bellman equation and can be found as the unique fixed point of the Bellman optimality operator $\mathcal{T}$ defined as $Q(x,a) = \mathcal{T}Q(x,a)$ with $\mathcal{T}Q(x,a) = \mathbb{E}_{\rho(x,a)}[R(x,a)] + \gamma \mathbb{E}_{P(\cdot|x,a)}[\max_{a'} Q(X', a')]$. To this end, TD learning consists of minimizing the squared temporal difference (TD) error, which is an estimation using the observed states, actions and rewards of the squared difference between a parameterized Q_θ and its update via the operator $\mathcal{T}Q_\theta$. In deep Q-learning, this is usually performed by considering an ϵ-greedy policy and a neural network to parameterize Q_θ.

In distributional RL (Bellemare et al., 2023), the idea is to consider the whole random variable $Z^\pi(x,a)$ rather than just its scalar expectation $Q^\pi(x,a)$. An analogous distributional Bellman equation can be derived between random variables. Denoting by $\bar{\eta}^\pi(x,a)$ the distribution of $Z^\pi(x,a)$, $\eta_1 * \eta_2$ the convolution between two distributions η_1 and η_2, and $(T_\gamma)_\#$ the pushforward operator through function $T_\gamma(z) = \gamma z$, we can write, $\bar{\eta}^\pi(x,a) = \rho(x,a) * \mathbb{E}_{P(\cdot|x,a),\pi}[(T_\gamma)_\# \bar{\eta}^\pi(X', A')]$, where the expectation corresponds to a mixture distribution over next states. This equation defines the so-called distributional Bellman operator denoted by $\mathcal{T}^\pi$ and so that $\bar{\eta}^\pi(x,a) = \mathcal{T}^\pi \bar{\eta}^\pi(x,a)$. See Proposition 2.17 and Figure 2.6 of Bellemare et al. (2023) for an illustration. Note that if $\bar{\eta}(x,a)$ (resp. $\mathcal{T}^\pi \bar{\eta}(x,a)$) is the probability density function of $Z(x,a)$ (resp. $\mathcal{T}^\pi Z(x,a)$), an equivalent random variable formulation, with $\overset{d}{=}$ meaning equality in distribution, is

$$\mathcal{T}^\pi Z(x,a) \overset{d}{=} R(x,a) + \gamma Z(X', A'),$$

where $X' \sim P(\cdot \mid x,a), A' \sim \pi(\cdot \mid x)$. This is referred to as the random-variable Bellman equation by Bellemare et al. (2023) (Proposition 2.16). The hope is then to find the return distribution as a fixed point of the distributional Bellman operator. The TD learning principle can be extended to differences between distributions leading to the minimization of quantities of the form $D(\mathcal{T}^\pi \bar{\eta}_1(x,a), \bar{\eta}_2(x,a))$ where D is a discrepancy or quasi-metric between distributions, equivalently denoted using random variables by $D(\mathcal{T}^\pi Z_1(x,a), Z_2(x,a))$. We will refer to D as a probability metric. Unfortunately, searching for a solution in the whole space of probability distributions is impossible. To use this approach, two main ingredients are required. We first need representations of distributions, rich enough to capture the return complexity and second, a choice of D for which a tractable fixed point algorithm can be implemented. This requires tractable evaluations of $D(\mathcal{T}^\pi Z_1(x,a), Z_2(x,a))$ and $\mathcal{T}^\pi$ being a contraction mapping with respect to $\overline{D}$ the supremum extension of D defined as $\overline{D}(\bar{\eta}_1, \bar{\eta}_2) = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} D(\bar{\eta}_1(x,a), \bar{\eta}_2(x,a))$, where $\bar{\eta}_1$ and $\bar{\eta}_2$ denote two collections of distributions, also named return functions by Bellemare et al. (2023) (Definition 4.21) .

2.1 Representations of the random return

For DRL, the goal is to define a good approximation (E, D) of the probability metric space to which $Z(x,a)$ belongs. Several ways to uniquely characterize a distribution are well known, such as the probability density function (PDF), the cumulative distribution function (CDF), the quantile function or the inverse CDF (QF), the characteristic function (Φ), etc. These representations are recalled in Appendix A. In the DRL literature, most choices are based on empirical or particles representations of the above, sometimes referred to as non parametric in statistics. In contrast, in this work, denoting by $\mathcal{F}$ a latent functional space giving the neural networks architectures, we propose to investigate a parametric choice using the PDF representation and $\mathcal{F} = \mathcal{M} = \cup_{K\in\mathbb{N}}\mathcal{M}_K$ where $\mathcal{M}_K$ is the set of univariate K-component Gaussian mixtures (GM) whose PDFs are of the form,

$$\eta(z) = \sum_{k=1}^{K} \pi_k \mathcal{N}(z; \mu_k, \sigma_k^2)$$

where $\mathcal{N}(\cdot; \mu_k, \sigma_k^2)$ or simply $\mathcal{N}(\mu_k, \sigma_k^2)$ denotes the Gaussian PDF with μ_k's (resp. σ_k^2's) the component means (resp. variances) and π_k's the components weights with $\pi_k \in [0, 1]$ and $\sum_{k=1}^{K} \pi_k = 1$.

2.2 Probability metrics

In DRL, probability metrics are used to provide a comparison of two random return distributions, generally in order to apply a TD learning principle. To compare distributions, in principle, all divergences are suitable candidates, but only a few have been used in DRL. The main used ones are the Kullback-Leibler (KL) divergence, the Cramer distance (also named energy distance) and the Wasserstein distance. The choice of metric may be particularly important as each metric offers different theoretical and practical properties. In DRL (Bellemare et al., 2017b), we have two first desirable properties to look for. First, we need to provide a contraction property for the distributional Bellman operator, *i.e.* to have $\alpha < 1$ such that for all possible return functions $\bar{\eta}_1$ and $\bar{\eta}_2$, $\overline{D}(\mathcal{T}^\pi \bar{\eta}_1, \mathcal{T}^\pi \bar{\eta}_2) \leq \alpha \overline{D}(\bar{\eta}_1, \bar{\eta}_2)$. This property is important as it implies the existence of a fixed point that can be reached by applying repeatedly the Bellman operator. To obtain such a property, D is often proved to be *ideal*, which is a sufficient condition assuming that D is p-convex (Theorem 4.25 of Bellemare et al. (2023) recalled in Appendix D.1). Being *ideal* means satisfying the two **(SI)** (sum-invariant) and **(S)** (scale-sensitive) properties below. For A, X, Y random variables, $\lambda \in (0, 1)$,

$$\begin{array}{ll} \textbf{(SI)} & A \perp (X, Y) \Longrightarrow D(A + X, A + Y) \leq D(X, Y) \\ \textbf{(S)} & \quad \exists c > 0, \quad D(\lambda X, \lambda Y) \leq \lambda^c D(X, Y). \end{array}$$

In addition, since minimizing $\overline{D}(\mathcal{T}^\pi \bar{\eta}_1, \bar{\eta}_2)$ is usually done using stochastic gradient-based algorithms, we also need to make sure that the numerically found optimum is the good one. For instance, it is proved by Bellemare et al. (2017b) that the Wasserstein distance has biased sample gradients which leads to convergence towards a wrong optima in practice according to the authors. Although it is possible to deal with biased stochastic gradient algorithms, see *e.g.* Rhee & Glynn (2015); Demidovich et al. (2023), a second important property of the metric is thus that it satisfies the following **(USG)** (Unbiased Sample Gradient) property (Bellemare et al., 2017b). Let η_θ be a probability distribution parameterized by θ and $\{X_m\}_{m\in[M]}$ a collection of M *i.i.d.* random variables distributed as $X \sim \eta_X$. Define $\hat{\eta}_M = \frac{1}{M}\sum_{m=1}^{M} \delta_{X_m}$ the empirical distribution of the $\{X_m\}_{m\in[M]}$, the (USG) property is satisfied if for all $M \in \mathbb{N}$,

$$\textbf{(USG)} \quad \mathbb{E}_{X_m \sim \eta_X} \left[\nabla_\theta D(\hat{\eta}_M, \eta_\theta)\right] = \nabla_\theta D(\eta_X, \eta_\theta).$$

As already mentioned, the Wasserstein distance does not satisfy the (USG) property.

In addition to theoretical properties, an important feature is the tractability of D to allow practical implementations. In this work, we are in particular interested in tractable D when comparing Gaussian mixtures. The following Table 1 summarizes, for a number of distances D, whether the four main characteristics important for practical DRL are satisfied. Details are given in the next sections. From the listed features, three metric-dependent properties can be derived: (1) a non-expansive projection, (2) a contracting Bellman operator, (3) a (USG) metric. While (1,2) are of theoretical nature, (3) is an algorithmic issue, that can be more easily addressed. As for QRDQN (Rowland et al., 2024), a solution is to find an (USG) metric with the same projection. Among metrics tractable for GM, the Cramer distance C (Section 6) and the mixture-specific Wasserstein distance MW_2 (Section 5) are thus the most promising ones as they satisfy (1,2) and experimental results in Section 8 show their practical effectiveness.

Table 1: Metrics properties. ✓ (resp.×) means satisfied (resp. unsatisfied), − means unknown. The KL, Cramer, MMD, Wasserstein metrics are studied by Bellemare et al. (2017b). The Jensen-Tsallis (JT) is proved to satisfy (USG) by Choi et al. (2019). MW_2 is a Wasserstein distance introduced by Delon & Desolneux (2020) but not investigated before in DRL. Properties showed in this paper are in red.

Metric properties				
Metric	SI	S	USG	GM Tractability
KL	-	×	✓	×
$\mathrm{MMD}_{k_{\mathrm{en}}}/\mathrm{C}$	✓	✓	✓	✓
$\mathrm{MMD}_{k_{\mathrm{rbf}}}$	-	×	✓	✓
$\mathrm{MMD}_{k\ \mathrm{lap}}$	-	-	✓	✓
Wasserstein	✓	✓	×	×
MW_2	✓	✓	×	✓
$\mathrm{JT}_{1,2}$	✓	×	✓	✓
$\mathrm{JT}_{x^2,2}$	×	✓	✓	✓

2.3 Analysis of GM dynamic programming

In practice, when choosing GM representations $\eta_\theta \in \mathcal{M}_K$, we also have to handle the fact that the result of the distributional Bellman operator on this class does not in general remain in the class. Practical implementations thus require a projection back on $\mathcal{M}_K$. When choosing a probability metric, it is then also important to check whether this projection is non expansive, so that the combination of the projection and the distributional Bellman operator is still a contraction with respect to the chosen metric. We further discuss and specify this aspect in the case of TD learning.

In principle, a contraction property of $\mathcal{T}^\pi$ deduced from (S) and (SI) is sufficient to guarantee the convergence of TD learning. However, a practical TD learning algorithm cannot represent the full return function $\mathcal{T}^\pi\eta$, which could be any return function. Hence, we need to take into account the approximation made at each iteration by the stochastic gradient descent. In this paper, we are only interested in the dynamic programming part of TD learning summarized in Algorithm 1, with a projection $\Pi_D\eta$ that is typically defined as a solution of $\min_{\eta'\in\mathcal{M}_K} D(\eta,\eta')$ for any $\eta \in \mathcal{M}$. This solution is not necessarily unique, so the projection needs to be parametrized with some parameter w^*. An example is provided in Section 5 and Appendix D.8. The full operator is then defined as $(\overline{\Pi_D^{w^*}}\bar{\eta})(x,a) = \Pi_D^{w^*}(\bar{\eta}(x,a))$ for all return function $\bar{\eta}$ and $(x,a) \in \mathcal{X}\times\mathcal{A}$. Using this definition, we can formulate our following main result below. As summarized in Table 1, only the MMD/Cramer (C), the Wasserstein and Mixture Wasserstein (MW_2) satisfy the (SI) and (S) conditions, which limits the hope to get convergence results for the others. Among the former ones, only the Cramer distance is (USG) but it is mentioned in Wiltzer et al. (2024) p.6 that the associated projection is not a non-expansion. In contrast, we show in what follows that the MW_2 metric is not in general (USG) (Theorem 5) but admits a non-expansive projection $\Pi_{MW_2}^{w^*}$ (Theorem 6). It follows Theorem 1 that states the projected Bellman operator is a contraction. All proofs are provided in the Appendix.

Algorithm 1: Mixture Dynamic Programming

```
Require: Mixture parameters estimates $(\theta(x) = (\theta_k(x))_{k=1}^K : x \in \mathcal{X})$, projection $\Pi_D$
for $x \in \mathcal{X}$ do
    Let $(x, R, X')$ define the random transition under policy $\pi$ with $a = \pi(x)$
    Update $\theta(x)$ to the mixture parameters corresponding to the distribution of $\Pi_D(R + \gamma\eta_{\theta(X')})$
Return: $\theta$
```

Theorem 1. *Assume that $\rho(x,a) \in \mathcal{M}$ for every state-action pair (x,a) (where $\rho(x,a)$ is the law of the reward $R(x,a)$), then $\overline{\Pi_{MW_2}^{w^*}}\mathcal{T}^\pi$ is a contraction mapping with respect to $\overline{MW_2}$.*

Thus applying Banach's fixed point theorem, Algorithm 1 converges towards a unique fixed point of $\overline{\Pi_{\mathrm{MW}_2}^{w^*}}\mathcal{T}^\pi$ if $R(x,a)$ is a Gaussian mixture. Unfortunately, as for quantile temporal difference (QTD) learning (Rowland et al., 2024), the sample-estimate of $\overline{\Pi^{w^*}_{\mathrm{MW}_2}\mathcal{T}^\pi}$ is biased (see Appendix D.9), so that we cannot straightforwardly deduce the convergence of our TD algorithm with stochastic gradient descent (SGD). However, we could hope to prove such a convergence by using techniques from Rowland et al. (2024) for QTD, with another metric that satisfies (USG) and for which the SGD converges to the same optima as MW_2.

3 Related work

Most methods (QR-DQN, IQN, FQF) use a quantile-based representation. The advantage of such an approach is that it is straightforward to derive the Monte-Carlo calculation of risk-aware policies with a distortion risk measure (Dabney et al., 2018a) using the functions largely studied in economics. Another advantage is that the change of variables, on the quantile function F^{-1}, $F^{-1}_{\gamma Z+r}(\tau) = \gamma F^{-1}_Z(\tau) + r$, induced by the distributional Bellman operator is easier to manipulate. Among quantile-based solutions, state-of-the-art performance is reached by the FQF method that builds a neural network supervising the choice of the sampled quantile fractions. However, it seems to fix an imaginary problem since it does not arise from the original RL problem but from the way we represent it and more particularly from the Monte-Carlo estimation of the 1D integral that defines the loss. In addition, the monotonicity of the represented quantile function is not guaranteed and the loss used (Huber variant of the Wasserstein loss) leads to biased gradient estimations. A more detailed discussion is given in Appendix B.

Therefore, it is interesting to search for alternative representations. In their paper, Choi et al. (2019) take an architecture inspired from Gaussian mixtures where the output of their neural network η_θ (named neural Gaussian mixture) is :

$$\eta_\theta(x,a) = \sum_{k=1}^{K} \pi_{k,\theta}(x,a)\mathcal{N}(\mu_{k,\theta}(x,a), \sigma^2_{k,\theta}(x,a)),$$

where $\pi_{k,\theta}(x,a), \mu_{k,\theta}(x,a), \sigma^2_{k,\theta}(x,a)$ have the same architecture as in DQN (Mnih et al., 2015). To compute the temporal difference, the authors use $D(X,Y) = \int_{\mathbb{R}} |\eta_X(z) - \eta_Y(z)|^2 dz$, where η_X, resp. η_Y, is the PDF of variable X, resp. Y. This metric is also used by Malekzadeh et al. (2023) in a different setting. However, it was shown recently by Zhang et al. (2024) that the DRL state-of-the-art could be outperformed using Gaussian neural mixtures with the Euclidean distance between the mixtures parameters instead. In this work, we investigate two other metrics detailed in the next sections.

4 Jensen-Tsallis divergence

The Jensen-Tsallis divergence is a generalization of the Jensen-Shannon divergence (Tsallis, 1988). In their work on DRL, Choi et al. (2019) refer to this divergence but they use it only in its simplest form $\mathrm{JT}_{1,2}$, which is the norm of the L_2 space. To better understand the issue with this metric, we consider a weighted version of it.

Definition 1. *Let $\omega : \mathbb{R} \to \mathbb{R}_+$ a measurable function. Let $p \in \mathbb{N}$, X and Y two random variables, whose PDFs, η_X and η_Y, are assumed to have their power p integrable with respect to the $\omega(z)dz$ measure. $JT_{\omega,p}$ is defined as,*

$$JT_{\omega,p}(X,Y) = \int_{\mathbb{R}} \omega(z)|\eta_X(z) - \eta_Y(z)|^p dz$$

and also denoted by $\|\eta_X - \eta_Y\|^p_{L_p(\omega)}$.

When $\omega = 1$, $p = 2$, we recover the Jensen-Tsallis metric used in Choi et al. (2019). This metric has the advantage to provide a closed-form formula for Gaussian mixtures. Let $Z_1 \sim \sum_{k=1}^{K_1} \pi_{1k}\ g_{1k}$ and $Z_2 \sim \sum_{k=1}^{K_2} \pi_{2k}\ g_{2k}$ two random variables distributed as Gaussian mixtures, with g_{ik} denoting a Gaussian PDF,

$$JT_{\omega,2}(Z_1,Z_2) = \sum_{k,\ell} \pi_{1k}\pi_{1\ell}\langle g_{1k}, g_{1\ell}\rangle_\omega + \sum_{k,\ell} \pi_{2k}\pi_{2\ell}\langle g_{2k}, g_{2\ell}\rangle_\omega - 2\sum_{k,\ell} \pi_{1k}\pi_{2\ell}\langle g_{1k}, g_{2\ell}\rangle_\omega,$$

where $\langle \cdot,\cdot\rangle_\omega$ denotes the weighted L_2 scalar product, which for two univariate Gaussian PDFs $g_1 = \mathcal{N}(\mu_1, \sigma_1^2)$ and $g_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ is closed-form, $\langle g_1, g_2\rangle_\omega = \mathcal{N}(\mu_1; \mu_2, \sigma_1^2 + \sigma_2^2)\ \mathbb{E}_G[\omega(G)]$, where $G \sim$

$\mathcal{N}\left(\frac{\mu_1\sigma_2^2+\mu_2\sigma_1^2}{\sigma_1^2+\sigma_2^2}, \frac{\sigma_1^2\sigma_2^2}{\sigma_1^2+\sigma_2^2}\right)$. Unfortunately, in its tractable form, the Jensen-Tsallis is not ideal. More specifically, the following result holds.

Theorem 2. *$JT_{1,2}$ is not ideal. Furthermore, if $\omega : \mathbb{R} \to \mathbb{R}_+$ is a measurable function such that $\omega(x) = 0 \implies x = 0$, we have, for A, X, Y random variables, with A independent on X and Y, and such that $\mathbb{E}[\omega(A)] < \infty$,*

$$JT_{\omega,2}(A+X, A+Y) \leq JT_{\omega,2}(X,Y)\mathbb{E}[\omega(A)]^2 \tag{1}$$

$$JT_{\omega,2}(\lambda X, \lambda Y) \leq \sup_{x\in\mathbb{R}} \frac{\omega(\lambda x)}{\lambda\omega(x)} JT_{\omega,2}(X,Y). \tag{2}$$

In particular, it can be deduced from (2) that $\mathrm{JT}_{1,2}$ does not satisfy (S). The distributional Bellman operator $\mathcal{T}^\pi$ is then not always a contraction mapping with respect to the associated metric. Moreover, it does not seem easy to correct the problem by taking another ω. For instance, $\mathrm{JT}_{x^2,2}$ satisfies (S) but does not always satisfy (SI) according to the previous result. See details in Appendix D.3.

The following result shows that if the reward is not noisy enough, then the distributional Bellman operator $\mathcal{T}^\pi$ cannot be a contraction mapping with $\overline{JT_{1,2}}$.

Theorem 3. *Let $\gamma \in (0,1)$ and X, Y two non-identically distributed random variables. There exists $\sigma_{max} > 0$ such that for any random variable A independent of X, Y, we have*

$$\mathbb{V}[A] \leq \sigma_{max} \Longrightarrow JT_{1,2}(A+\gamma X, A+\gamma Y) > JT_{1,2}(X,Y),$$

where $\mathbb{V}[A]$ denotes the variance of A.

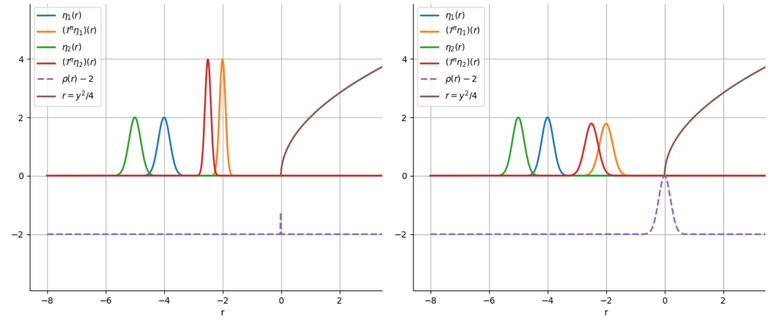
Hence, if in a deterministic (or sufficiently low-noise) setting and all states are accessible, $\mathcal{T}^\pi$ is not a contraction mapping with respect to $\overline{JT_{1,2}}$. A simple illustration of the result can be given in a simplified Gaussian setting. Consider a policy π that maps any state x to a single action a_0 and a transition that leaves the state invariant, *i.e.* $\pi(\cdot|x) = \delta_{a_0}(\cdot)$ and $P(\cdot|x, a_0) = \delta_x(\cdot)$. Consider then a reward with a Gaussian distribution $\rho(x, a_0) = \mathcal{N}(0, \sigma_R^2)$. For two Gaussian distributions $\eta_1(x, a_0) = \mathcal{N}(\mu_1, \sigma_Z^2)$ and $\eta_2(x, a_0) = \mathcal{N}(\mu_2, \sigma_Z^2)$, it follows that $\mathcal{T}^\pi\eta_i(x, a_0) = \mathcal{N}(\gamma\mu_i, \gamma^2\sigma_Z^2 + \sigma_R^2)$ for any γ and $i = 1, 2$. For Gaussian mixtures and single Gaussian distributions, the $\mathrm{JT}_{1,2}$ metric is available in closed-form, so that $JT_{1,2}(\mathcal{T}^\pi\eta_1, \mathcal{T}^\pi\eta_2) = \frac{1}{\sqrt{\pi}\sqrt{\gamma^2\sigma_Z^2+\sigma_R^2}}\left(1 - \exp\left(-\frac{\gamma^2(\mu_1-\mu_2)^2}{4(\gamma^2\sigma_Z^2+\sigma_R^2)}\right)\right)$ and $JT_{1,2}(\eta_1, \eta_2) = \frac{1}{\sqrt{\pi}\sqrt{\sigma_Z^2}}\left(1 - \exp\left(-\frac{(\mu_1-\mu_2)^2}{4\sigma_Z^2}\right)\right)$. It is then easy to see that if $\gamma^2 < 1$ and $\sigma_Z^2 > \gamma^2\sigma_Z^2 + \sigma_R^2$, then $JT_{1,2}(\mathcal{T}^\pi\eta_1, \mathcal{T}^\pi\eta_2) > JT_{1,2}(\eta_1, \eta_2)$. And this occurs as soon as $\gamma \in (0,1)$ and σ_R is taken low enough. Conversely, when σ_R increases $JT_{1,2}(\mathcal{T}^\pi\eta_1, \mathcal{T}^\pi\eta_2)$ decreases and can go below $JT_{1,2}(\eta_1, \eta_2)$. A graphical illustration is given in Figure 1, with $a_0 = 0$, $\gamma = 0.5$ and $\sigma_R = 0.002$ or $\sigma_R = 0.2$. For low values of σ_R, the effect of the $\mathcal{T}^\pi$ is to decrease variances and thus increase the separation between the two Gaussians and consequently their $\mathrm{JT}_{1,2}$ distance (Figure 1-(a)). In contrast for higher σ_R (Figure 1-(b)), $\mathcal{T}^\pi\eta_1$ and $\mathcal{T}^\pi\eta_2$ are more intertwined and the distance decreases. To address this contraction issue with the $\mathrm{JT}_{1,2}$ metric, we propose studying other metrics that seem more promising.

5 Mixture-Wasserstein distance

Delon & Desolneux (2020) have introduced a distance specifically designed for Gaussian mixtures based on the Wasserstein distance. In an optimal transport context, by restricting the possible coupling measures (*i.e.*, the optimal transport plan) to a Gaussian mixture, they propose a discrete formulation for this distance. This makes it tractable while in general using the standard Wasserstein distance between mixtures is problematic. Delon & Desolneux (2020) refer to the proposed new distance as MW_2, for *Mixture Wasserstein*. The MW_2 definition makes use of the tractability of the Wasserstein distance between two Gaussians for a quadratic cost. The standard quadratic cost Wasserstein distance between two univariate Gaussian PDFs $g_1 = \mathcal{N}(\mu_1, \sigma_1^2)$, $g_2 = \mathcal{N}(\mu_2, \sigma_2^2)$ is,

$$\mathrm{W}_2^2(g_1, g_2) = (\mu_1 - \mu_2)^2 + (\sigma_1 - \sigma_2)^2. \tag{3}$$

Section 4 of Delon & Desolneux (2020) shows that the MW_2 distance between two mixtures can be computed by solving a discrete transport problem.



(a) Low-noise ($\sigma_R = 0.002$)

(b) Higher noise ($\sigma_R = 0.2$)

Figure 1: $\mathrm{JT}_{1,2}$ contraction in low (a) and higher (b) noise. The purple curve shows the $(1-\exp(-y^2/4)) \approx y^2/4$ term in the $\mathrm{JT}_{1,2}$ expression and illustrates how the distance increases with the separation (y).

Definition 2. *Let $\eta_1 = \sum_{k=1}^{K_1} \pi_{1k}\ g_{1k}$ and by $\eta_2 = \sum_{k=1}^{K_2} \pi_{2k}\ g_{2k}$ be two Gaussian mixtures. Then, the mixture Wasserstein distance MW_2 is defined as,*

$$MW_2^2(\eta_1, \eta_2) = \min_{w \in \Pi(\pi_1, \pi_2)} \sum_{k,\ell} w_{k\ell}\ \mathrm{W}_2^2(g_{1k}, g_{2\ell}) \tag{4}$$

where π_1 and π_2 are the discrete distributions on the simplex defined by the respective weights of the mixtures and $\Pi(\pi_1, \pi_2)$ is the set of discrete joint distributions $w = (w_{k\ell} \in [0,1], k \in [K_1], \ell \in [K_2])$, whose marginals are π_1 and π_2.

Finding the minimizer w^* of (4) boils down to solving a simple discrete optimal transport problem, where the entries of the $K_1 \times K_2$ dimensional cost matrix are the $\mathrm{W}_2^2(g_{1k}, g_{2\ell})$ quantities. As implicitly suggested above, MW_2 is indeed a distance on the space of Gaussian mixtures; see Delon & Desolneux (2020). In particular, for two Gaussian mixtures η_1 and η_2, MW_2 satisfies the equality property according to which $\mathrm{MW}_2(\eta_1, \eta_2) = 0$ implies that $\eta_1 = \eta_2$. Expression (4) is interesting as when using (3), it can be favorably compared to the Euclidean distance used in Zhang et al. (2024). The latter compares parameters of two Gaussian mixtures, which need to have the same number of components in a prescribed order for the comparison to make sense. The MW_2 distance instead is a generalization, that can be computed between any mixtures without requiring manual alignment, thanks to the transport map. In addition, when considering mixtures η_1 and η_2 of components ℓ_{ik} in a distributions family $\mathcal{L}$, we can define the following generalization,

$$\mathrm{MW}_{D,\mathcal{L}}^p(\eta_1, \eta_2) = \min_{w \in \Pi(\pi_1, \pi_2)} \sum_{k,l} w_{kl}\ D^p(\ell_{1k}, \ell_{2l}).$$

When D is the Euclidean distance and $p = 2$, such a generalization is discussed in Section 4.6 of Delon & Desolneux (2020). For instance, mixtures of elliptical distributions satisfy the required properties, in particular when considering the easier case of univariate mixtures. In what follows, we will thus assume that $\mathrm{MW}_{D,\mathcal{L}}^p$ is a metric. Details on results that are still valid if $\mathrm{MW}_{D,\mathcal{L}}^p$ is a quasi-metric are given in Appendix D.6. We can show the following Lemma.

Lemma 1. *Assume that D is ideal and $\mathcal{L}$ is stable by scaling and summing, then $MW_{D,\mathcal{L}}^p$ is ideal.*

It follows from Lemma 1 and Theorem 4.25 in Bellemare et al. (2023) that $\mathrm{MW}_{D,\mathcal{L}}^p$ can make $\mathcal{T}^\pi$ a contraction while maintaining tractability. More specifically, we prove the following result.

Theorem 4. *Assume that D is ideal, $\mathcal{L}$ is stable by scaling and summing and $\rho(x,a) \in \mathcal{L}$ for every state-action pair (x,a) (where $\rho(x,a)$ is the law of the reward $R(x,a)$), then $\mathcal{T}^\pi$ is a contraction mapping with respect to $MW_{D,\mathcal{L}}^p$.*

$\mathrm{MW}_{D,\mathcal{L}}^p$ variants and in particular MW_2^2 provide then new interesting alternative metrics for neural mixtures. Unfortunately, the MW_2^2 distance is not satisfying in general the (USG) property and does not have unbiased sample gradients. Indeed, note that MW_2 and W_2 coincide on Dirac mixtures and W_2 does not have unbiased sample gradients (see Bellemare et al. (2017b)) so that it is impossible for MW_2 to satisfy (USG). The following theorem gives the exact formulation of the bias.

Theorem 5. *Let $\hat{\eta}_M = \frac{1}{M}\sum_{m=1}^{M} \delta_{X_m}$ be the empirical distribution of M i.i.d. $\{X_m\}_{m\in[M]}$ from $\eta_X = \sum_{\ell=1}^{K^X} \pi_{\ell,X}\mathcal{N}(\mu_{\ell,X}, \sigma^2_{\ell,X})$. Define $\eta_\theta = \sum_{k=1}^{K} \pi_{k,\theta}\mathcal{N}(\mu_{k,\theta}, \sigma^2_{k,\theta})$ parameterized by θ then,*

$$\mathbb{E}_{X_m\sim\eta_X}\left[\nabla_\theta MW_2^2(\hat{\eta}_M, \eta_\theta)\right] - \nabla_\theta MW_2^2(\eta_X, \eta_\theta) =$$

$$2\nabla_\theta\left(\sum_{k,\ell} \tilde{w}^*_{k\ell}(\mu_{\ell,X}\mu_{k,\theta} + \sigma_{\ell,X}\sigma_{k,\theta})\right) - 2\nabla_\theta\left(\sum_{k} \mu_{k,\theta}\mathbb{E}[\sum_{m} w^*_{km}(\bar{X})X_m]\right)$$

where $\tilde{w}^*_{k\ell} \in \Pi(\pi_\theta, \pi_X)$ is the optimal coupling defining $MW_2^2(\eta_X, \eta_\theta)$ and w^*_{km} that of $MW_2^2(\hat{\eta}_M, \eta_\theta)$.

The right-hand side above is in general non zero. If all variances in η_θ are constant, *i.e.* $\sigma_{k,\theta} = \sigma$, then $\nabla_\theta\left(\sum_{k,\ell} \tilde{w}^*_{k,\ell}\sigma_{k,\theta}\sigma_{\ell,X}\right) = \sigma\nabla_\theta\left(\sum_\ell \pi_{\ell,X}\sigma_{\ell,X}\right) = 0$. This condition would be acceptable in practice as the resulting GM remain flexible models but it is not enough to cancel the right-hand side. However, (USG) is trivially satisfied if η_θ has only one component ($K = 1$).

Nevertheless, MW_2^2 remains an interesting metric. Theorem 6 below shows the existence of a non-expansive projection. Combined with Theorem 4, it guaranties that the corresponding projected Bellman operator is a contraction mapping with respect to $\overline{\mathrm{MW}_2}$ as announced in Theorem 1. As discussed in Section 2.3, this is an important feature to analyse the behavior of a TD algorithm.

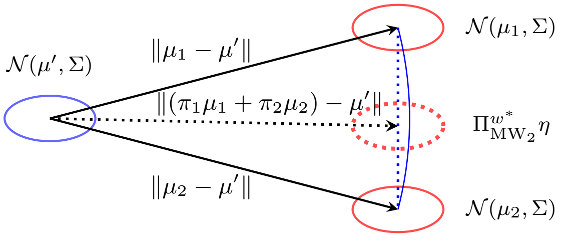
Theorem 6. *Let $K \in \mathbb{N}$. There exists a set W^*_K of functions $w^* : \mathcal{M} \to [0,1]^{\mathbb{N}\times K}$ verifying for all $L \in \mathbb{N}$, $\ell \in [L]$, $\sum_{k=1}^{K} w^*_{\ell k}(\eta) = \pi_\ell$, where $\eta = \sum_{\ell=1}^{L} \pi_\ell\mathcal{N}(\mu_\ell, \sigma^2_\ell)$, such that the MW_2 projections are characterized by $w^* \in W^*_K$ and for all $\eta \in \mathcal{M}$, the projection is defined by*

$$\Pi^{w^*}_{MW_2}\eta = \sum_{k=1}^{K} \tilde{\pi}_k(\eta)\mathcal{N}(\tilde{\mu}_k(\eta), \tilde{\sigma}_k(\eta)^2),$$

with $\tilde{\pi}_k(\eta) = \sum_\ell w^*_{\ell k}(\eta)$, $\tilde{\mu}_k(\eta) = \sum_\ell \frac{w^*_{\ell k}(\eta)\mu_\ell}{\tilde{\pi}_k(\eta)}$ and $\tilde{\sigma}_k(\eta) = \sum_\ell \frac{w^*_{\ell k}(\eta)}{\tilde{\pi}_k(\eta)}\sigma_\ell$.

Moreover, $\Pi^{w^*}_{MW_2}$ is a non-expansion with respect to MW_2 for all $w^* \in W^*_K$.

As an example, if $\eta' = \mathcal{N}(\mu', \sigma^2)$ and $\eta = \pi_1\mathcal{N}(\mu_1, \sigma^2) + \pi_2\mathcal{N}(\mu_2, \sigma^2)$ then $\Pi^{w^*}_{\mathrm{MW}_2}\eta = \mathcal{N}(\pi_1\mu_1 + \pi_2\mu_2, \sigma^2)$, $\mathrm{MW}_2^2(\eta, \eta') = \pi_1(\mu_1 - \mu')^2 + \pi_2(\mu_2 - \mu')^2$ and $\mathrm{MW}_2^2(\Pi^{w^*}_{\mathrm{MW}_2}\eta, \eta') = (\pi_1\mu_1 + \pi_2\mu_2 - \mu')^2$. It follows from the Cauchy–Schwarz inequality that $\mathrm{MW}_2^2(\Pi^{w^*}_{\mathrm{MW}_2}\eta, \eta') \leq \mathrm{MW}_2^2(\eta, \eta')$. Figure 2 gives a graphical illustration, where the non-expansion property is more clearly visualized in 2D, below in the case $||\mu_1 - \mu'|| = ||\mu_2 - \mu'||$.



$\mathcal{N}(\mu', \Sigma)$
$\|\mu_1 - \mu'\|$
$\mathcal{N}(\mu_1, \Sigma)$
$\|(\pi_1\mu_1 + \pi_2\mu_2) - \mu'\|$
$\Pi^{w^*}_{\mathrm{MW}_2}\eta$
$\|\mu_2 - \mu'\|$
$\mathcal{N}(\mu_2, \Sigma)$

Figure 2: Illustration of the non-expansion $\Pi^{w^*}_{\mathrm{MW}_2}$ (Theorem 6) for $K = 1$ and $L = 2$.

6 Maximum Mean Discrepancy

In their work, Nguyen-Tang et al. (2020) use MMD for Q-learning in DRL with particles and have demonstrated competitive performance. Different MMD exist depending on a choice of kernel.

Definition 3. *Let $k : (\mathbb{R}^d)^2 \to \mathbb{R}$ a kernel. Let P, Q be two distributions and $X, \tilde{X}$ (resp. $Y, \tilde{Y}$) two independent variables following P (resp. Q), with X also independent of Y. The squared MMD (Maximum Mean Discrepancy) between P and Q is defined as*

$$MMD_k^2(P, Q) = \mathbb{E}[k(X, \tilde{X})] + \mathbb{E}[k(Y, \tilde{Y})] - 2\mathbb{E}[k(X, Y)].$$

If k is a reproducing kernel, the metric MMD_k is equal to $\|\mu_P - \mu_Q\|_{\mathcal{K}}$ where $\mathcal{K}$ is the RKHS associated to k and μ_P (resp. μ_Q) is the mean on this RKHS of P (resp. Q). In this work, we only consider translation-invariant kernels and more particularly, the Laplacian kernel $k_{\gamma_0,\text{lap}}(x, y) = e^{-\gamma_0\|x-y\|_2}$, the Gaussian kernel $k_{\gamma_1,\text{rbf}}(x, y) = e^{-0.5\gamma_1^{-2}\|x-y\|_2^2}$ and the energy kernel $k_{\text{en}}(x, y) = -\|x - y\|_2$.

The reason why Gaussian kernels perform better than energy kernels, called unrectified kernels by Nguyen-Tang et al. (2020), is still unclear in the literature. Indeed, it is shown by Killingberg & Langseth (2023) that the so-called multiquadric kernel seems better than the Gaussian kernel on a theoretical point of view and less sensitive to hyperparameters. In this work, we propose to shed another light on these kernels comparison by using them with our GM representations. We show in Theorem 10 in Appendix D.11 that the above kernels all lead to tractable formulas for GM.

Regarding the use of MMD with Gaussian mixtures, Nam et al. (2021) have shown the good performance of the Cramer distance, which is equivalent to the energy distance, as recalled in Theorem 11 in Appendix D.12. However, they use a policy-gradient algorithm and no real assessment has been made for TD algorithms, although they have been considered in Zhang (2023). The MMD metric always satisfies (USG) (Bellemare et al., 2017b) so all the good properties of $\text{JT}_{w,2}$ are recovered but it also makes $\mathcal{T}^\pi$ a contraction mapping for the energy kernel. In Nguyen-Tang et al. (2020), the authors report that the energy kernel (called unrectified kernel in their paper) does not seem to give promising results. In contrast, they show that a mixture of Gaussian kernels give better results in practice although the MMD with Gaussian kernel suffers from the same problem as $\text{JT}_{1,2}$. They explain this with a moment-matching-like property of $\text{MMD}_{k_{\gamma_1,\text{rbf}}}$ that should also be valid in our case with Gaussian mixtures. However, we show in our experiments that the energy kernel leads to better results, suggesting that the results obtained in Nguyen-Tang et al. (2020) might be due to their choice of representations using particles.

7 Generalization to multidimensional rewards

Another good feature of our proposed GM setting is its natural generalization to higher dimensions. Although a full empirical analysis of multi-objective RL approaches would go beyond the scope of this paper, this section provides some theoretical discussion and first steps towards multidimensional rewards. Generalizing DRL algorithms to the multi-objective case requires to consider multidimensional rewards; see Appendix C.3 for details. This has been considered before but only for independent dimensions (Zhou et al., 2021) or with particles (Wiltzer et al., 2024). However, accounting for dependence between objectives allows to keep track of more information and the use of multivariate particles (or quantiles) seems unsatisfying as quantiles do not generalize satisfyingly in dimension greater than one. In contrast, multivariate Gaussian mixtures provide a natural way to define a TD algorithm for multidimensional rewards. We can easily generalize our previous results in the multidimensional case except for the projection associated to MW_2. The MW_2 as defined by Delon & Desolneux (2020), corresponds to our $\text{MW}^p_{D,\mathcal{L}}$, introduced in Section 5, with $D = W_2$ the 2-Wasserstein distance. This distance between two multivariate Gaussian distributions in dimension d, $g_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ and $g_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$ is, $\text{W}_2^2(g_1, g_2) = \|\mu_1 - \mu_2\|^2 + Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{1/2}\Sigma_2\Sigma_1^{1/2})^{1/2})$, and there is no obvious way to show that the associated projection is a non-expansion. As an alternative, we propose to make use of our $\text{MW}^p_{D,\mathcal{L}}$ generalization to define a new multivariate metric $\text{MW}^2_{F,\mathcal{M}^d}$ where $\mathcal{M}^d$ is the set of d-variate Gaussian mixtures and, $F^2(g_1, g_2) = \|\mu_1 - \mu_2\|^2 + Tr((\Sigma_1^{1/2} - \Sigma_2^{1/2})^2)$. For two multivariate GM in dimension d, $\eta_1 = \sum_{k=1}^{K_1} \pi_{1k}\mathcal{N}_d(\mu_{1k}, \Sigma_{1k})$ and $\eta_2 = \sum_{k=1}^{K_2} \pi_{2k}\mathcal{N}_d(\mu_{2k}, \Sigma_{2k})$, we define,

$$\text{MW}^2_{F,\mathcal{M}^d}(\eta_1, \eta_2) = \min_{w \in \Pi(\pi_1,\pi_2)} \sum_{k,\ell} w_{k\ell} \left(\|\mu_{1k} - \mu_{2\ell}\|^2 + Tr((\Sigma_{1k}^{1/2} - \Sigma_{2\ell}^{1/2})^2)\right).$$

It is possible to check that $\mathrm{MW}^2_{F,\mathcal{M}^d}$ is ideal using Lemma 1 and to derive the same kind of non-expansive projection as for MW_2 (details are in Appendix D.10). Thus, we obtain the following result generalizing Theorem 1 to the multivariate case.

Theorem 7. *Assume that $\rho(x, a) \in \mathcal{M}^d$ for every state-action pair (x, a) (where $\rho(x, a)$ is the law of the reward $R(x, a)$), then $\overline{\Pi^{w^*}_{MW^2_{F,\mathcal{M}^d}}}\mathcal{T}^\pi$ is a contraction mapping with respect to $\overline{MW^2_{F,\mathcal{M}^d}}$.*

8 Experiments

Considering a standard TD learning framework, we first compare our proposed algorithm with Gaussian mixtures and different metrics, by running our agent on a selected subset of Atari games referred to as Atari-5. Atari-5 is a subset of 5 representative Atari games exhibited in the study of Aitchison et al. (2023) based on which a global normalized score can be computed using weights. These games are by default deterministic. We use the same standard architecture as in DQN (Mnih et al., 2015). We then also illustrate the different metrics behavior on the previous Atari games, with a modified MDP by adding sticky actions with probability 0.25, bringing stochasticity. Details on the hyperparameters used in these experiments are provided in the Appendix. The compared metrics are MW_2, $\mathrm{JT}_{1,2}$ and two MMDs. As in Nguyen-Tang et al. (2020), we consider the kernel built as the following sum of Gaussian kernel $k_{\text{mix rbf}}(x, y) = \sum_{i=1}^{10} k_{\sqrt{i},\text{rbf}}(x, y)$ and the energy kernel. The results for the Gaussian kernel are not shown as it can be seen as a simpler version of $k_{\text{mix rbf}}$ and we observed that the Laplacian kernel behaved similarly to the Gaussian one. Results showing average normalized scores and their standard deviations (over 5 runs), using the weights given in Aitchison et al. (2023), are reported in Table 2. $\mathrm{MMD}_{\mathrm{k_{en}}}$, $\mathrm{MMD}_{\mathrm{k_{mixrbf}}}$, MW_2, perform better than the $\mathrm{JT}_{1,2}$ metric. Interestingly, no environment shows a real advantage of $\mathrm{MMD}_{\mathrm{k_{mixrbf}}}$ over $\mathrm{MMD}_{\mathrm{k_{en}}}$, which contrasts with the results of Nguyen-Tang et al. (2020). In particular, in the deterministic setting, the mean normalized score is better for $\mathrm{MMD}_{\mathrm{k_{en}}}$. The results also illustrate the superiority of our proposed metrics compared to $\mathrm{JT}_{1,2}$, whenever the environment is not stochastic enough, as suggested by Theorem 3.

Table 2: Metrics comparison on Atari-5. Average final normalized scores and standard deviations (over 5 runs) using the weights given in Aitchison et al. (2023) (best in bold characters).

Environment	$\mathbf{JT}_{1,2}$	$\mathbf{MW}_2$	$\mathbf{MMD}_{\mathrm{k_{en}}}$	$\mathbf{MMD}_{\mathrm{k_{mixrbf}}}$
Stochastic	37.24 ± 6.02	43.38 ± 4.72	46.84 ± 4.31	$\mathbf{57.66 \pm 4.47}$
Deterministic	34.06 ± 6.89	39.90 ± 14.77	$\mathbf{57.82 \pm 18.67}$	54.07 ± 8.54

Figure 3 shows in the deterministic (first row) and stochastic environment (second row), moving average returns over 30 millions of training frames for each game and the resulting normalized score. For games *Qbert* and *Name This Game*, the gap between the $\mathrm{JT}_{1,2}$ and other metrics learning curves is clearly reduced in the stochastic case. This is also observed in the normalized results (last column), comforting our theoretical statement that the noisier the environment, the more contractive $\mathcal{T}^\pi$ becomes with respect to $\mathrm{JT}_{1,2}$.

We then perform an ablation study, varying the number of mixture components $K \in \{1, 3, 5, 16\}$ in the deterministic case with the Atari-3 subset of 3 games proposed in Aitchison et al. (2023). Results for metric MW_2 are reported in Appendix Table 4. Surprisingly, a too high $K = 16$ seems to degrade results. However, Appendix Figure 5, *e.g.* for the *Battle Zone* game, suggests that the $K = 16$ learning curve might catch up with the other curves with more training frames. The similarly performance for $K \in \{1, 3, 5\}$ is also probably due to a too small number of training frames. In contrast, in different environments, such as that of the *Asterix* game, the $K = 5$ results are significantly better than the $K = 1$ case with more training frames (200 millions). Experiments made on a set of 3 other Atari games (see Figure 6 in Appendix) corroborate this analysis when the number of training frames is increased to 175 millions.

Finally, we observe that the metrics we propose do not introduce much additional computation cost except for MW_2 (see Tables 5 and 6 in Appendix). In the MW_2 metric case, most of the computation overhead is due to the optimal transport (EMD) solver used internally. This computation overhead may be reduced by reducing K or optimizing the implementation using recent developments in applied optimal transport.

Our code is implemented in Python and is available at `https://gitlab.inria.fr/mantonet/gm-drl`.

9 Conclusion

We showed that mixtures were interesting models to represent distributions in DRL. Their tractability and expressiveness allow to consider various metrics and result in implementations requiring less parameters than other algorithms. In a standard TD learning setting, we illustrated that the simple $\mathrm{JT}_{1,2}$ metric may not be suitable and proposed alternative Wasserstein-like and MMD metrics, with better theoretical and empirical properties. In particular, the proposed Mixture-Wasserstein metric showed both tractability and promising performance. To fully justify its use in stochastic gradient-based methods, it would be useful to study whether it is possible to handle its biased sample gradients. We also proposed an extended formulation $\mathrm{JT}_{\omega,p}$ of the Jensen-Tsallis distance with an additional weight term ω, which provides more flexibility and could be further exploited. Additionally, as briefly discussed in Section 7, the MW setting provides a promising approach to multi-objective DRL but a more complete study was out of the scope of this paper. In Appendix Section E, we also presented another advantage of mixtures, which is their particular adaptability to an alternative way to solve the distributional Bellman fixed point using a stochastic approximation principle. This alternative opens the way to the design of new efficient DRL procedures whose investigation is left for future work.

Acknowledgments

The authors were granted access to the HPC resources of IDRIS under the allocation 2024-A0171015651 made by GENCI.

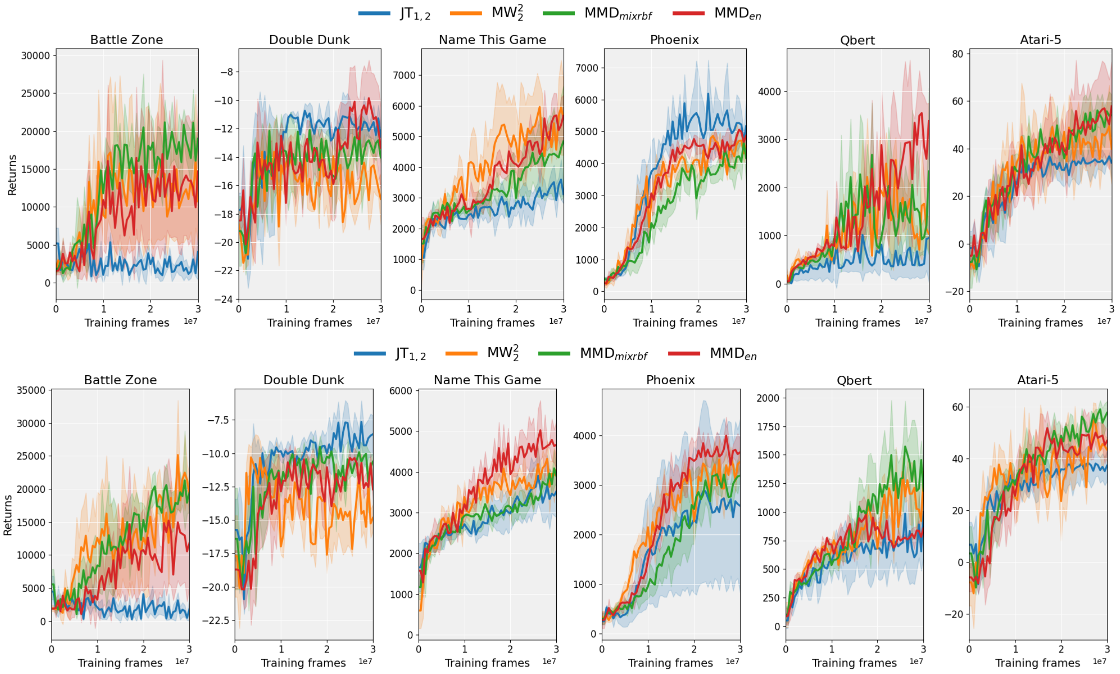


Figure 3: Comparison of MW_2(orange), $\mathrm{JT}_{1,2}$(blue), $\mathrm{MMD}_{k_{en}}$(red) and $\mathrm{MMD}_{k_{mixrbf}}$(green) metrics for Atari-5 games in a deterministic (first row) and stochastic (second row) environment. Moving average return for each game and normalized score, over 50 episodes, with respect to the number of training frames. Curves are averaged over 5 runs with shaded areas representing standard deviations.

References

Matthew Aitchison, Penny Sweetser, and Marcus Hutter. Atari-5: Distilling the arcade learning environment down to five games. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 421–438. PMLR, 23–29 Jul 2023.

Marc G. Bellemare, Will Dabney, and Rémi Munos. A distributional perspective on reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70, pp. 449–458. PMLR, 06–11 Aug 2017a.

Marc G. Bellemare, Ivo Danihelka, Will Dabney, Shakir Mohamed, Balaji Lakshminarayanan, Stephan Hoyer, and Rémi Munos. The Cramer Distance as a Solution to Biased Wasserstein Gradients, 2017b. URL https://arxiv.org/abs/1705.10743.

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

Fabio Bellini and Elena Di Bernardino. Risk management with expectiles. *The European Journal of Finance*, 23(6):487–506, 2017.

Vladimir Bogachev. *Measure theory*, volume I of *Theorem 3.9.4*. Springer, 2007.

Xin-Qiang Cai, Pushi Zhang, Li Zhao, Jiang Bian, Masashi Sugiyama, and Ashley Llorens. Distributional pareto-optimal multi-objective reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 36, pp. 15593–15613, 2023.

Pablo Samuel Castro, Subhodeep Moitra, Carles Gelada, Saurabh Kumar, and Marc G. Bellemare. Dopamine: A Research Framework for Deep Reinforcement Learning, 2018. URL http://arxiv.org/abs/1812.06110.

Arthur Charpentier. An introduction to multivariate and dynamic risk measures, 2018. URL https://hal.science/hal-01831481/.

Zaiwei Chen, Siva Theja Magulur, Sanjay Shakkottai, and Karthikeyan Shanmugam. Finite-Sample Analysis of Contractive Stochastic Approximation Using Smooth Convex Envelopes. In *Advances in Neural Information Processing Systems*, volume 33, pp. 8223–8234, 2020.

Yunho Choi, Kyungjae Lee, and Songhwai Oh. Distributional Deep Reinforcement Learning with a Mixture of Gaussians. In *The 2019 International Conference on Robotics and Automation (ICRA)*, 2019.

Will Dabney, Georg Ostrovski, David Silver, and Remi Munos. Implicit quantile networks for distributional reinforcement learning. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80, pp. 1096–1105. PMLR, 10–15 Jul 2018a.

Will Dabney, Mark Rowland, Marc Bellemare, and Rémi Munos. Distributional reinforcement learning with quantile regression. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018b.

Julie Delon and Agnès Desolneux. A Wasserstein-type distance in the space of Gaussian Mixture Models. *SIAM Journal on Imaging Sciences*, 13(2):936–970, 2020.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A Guide Through the Zoo of Biased SGD. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023.

Remi Flamary, Nicolas Courty, Alexandre Gramfort, Mokhtar Z. Alaya, Aurelie Boisbunon, Stanislas Chambon, Laetitia Chapel, Adrien Corenflos, Kilian Fatras, Nemo Fournier, Leo Gautheron, Nathalie T.H. Gayraud, Hicham Janati, Alain Rakotomamonjy, Ievgen Redko, Antoine Rolet, Antony Schutz, Vivien Seguy, Danica J. Sutherland, Romain Tavenard, Alexander Tong, and Titouan Vayer. POT: Python Optimal Transport. *Journal of Machine Learning Research*, 22(78):1–8, 2021.

Conor F. Hayes, Roxana Radulescu, Eugenio Bargiacchi, Johan Källström, Matthew Macfarlane, Mathieu Reymond, Timothy Verstraeten, Luisa M. Zintgraf, Richard Dazeley, Fredrik Heintz, Enda Howley, Athirai A. Irissappane, Patrick Mannion, Ann Nowé, Gabriel de Oliveira Ramos, Marcello Restelli, Peter Vamplew, and Diederik M. Roijers. A practical guide to multi-objective reinforcement learning and planning. *CoRR*, abs/2103.09568, 2021. URL `https://arxiv.org/abs/2103.09568`.

Elia Kaufmann, Leonard Bauersfeld, Antonio Loquercio, Matthias Müller, Vladlen Koltun, and Davide Scaramuzza. Champion-level drone racing using deep reinforcement learning. *Nature*, 620:982–987, 2023.

Ludvig Killingberg and Helge Langseth. The Multiquadric Kernel for Moment-Matching Distributional Reinforcement Learning. *Transactions on Machine Learning Research*, pp. 1–17, 2023.

Zhijun Li, Ting Zhao, Fei Chen, Yingbai Hu, Chun-Yi Su, and Toshio Fukuda. Reinforcement learning of manipulation and grasping using dynamical movement primitives for a humanoidlike mobile manipulator. *IEEE/ASME Transactions on Mechatronics*, 23(1):121–131, 2018.

Elliott H Lieb. Convex trace functions and the Wigner-Yanase-Dyson conjecture. *Advances in Mathematics*, 11(3):267–288, 1973.

Parvin Malekzadeh, Ming Hou, and Konstantinos N. Plataniotis. A unified uncertainty-aware exploration: Combining epistemic and aleatory uncertainty. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5, 2023.

John Martin, Michal Lyskawinski, Xiaohu Li, and Brendan Englot. Stochastically dominant distributional reinforcement learning. In *Proceedings of the 37th International Conference on Machine Learning*. PMLR, 2020.

Borislav Mavrin, Hengshuai Yao, Linglong Kong, Kaiwen Wu, and Yaoliang Yu. Distributional reinforcement learning for efficient exploration. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 4424–4434. PMLR, 09–15 Jun 2019.

Volodymyr Mnih, Koray Kavukcuoglu, David Silver, Andrei A. Rusu, Joel Veness, Marc G. Bellemare, Alex Graves, Martin Riedmiller, Andreas K. Fidjeland, Georg Ostrovski, Stig Petersen, Charles Beattie, Amir Sadik, Ioannis Antonoglou, Helen King, Dharshan Kumaran, Daan Wierstra, Shane Legg, and Demis Hassabis. Human-level control through deep reinforcement learning. *Nature*, 518(7540):529–533, 2015.

Daniel W Nam, Younghoon Kim, and Chan Y Park. GMAC: A Distributional Perspective on Actor-Critic Framework. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139, pp. 7927–7936. PMLR, 18–24 Jul 2021.

T.T. Nguyen, F. Chamroukhi, H.D. Nguyen, and G. J. McLachlan. Approximation of probability density functions via location-scale finite mixtures in Lebesgue spaces. *Communications in Statistics - Theory and Methods*, 52(14):5048–5059, 2023.

Thanh Nguyen-Tang, Sunil Gupta, and Svetha Venkatesh. Distributional Reinforcement Learning via Moment Matching. In *Proceedings of the AAAI Conference on Artificial Intelligence*. AAAI, 2020.

Chang-han Rhee and Peter W Glynn. Unbiased estimation with square root convergence for SDE models. *Operations Research*, 63(5):1026–1043, 2015.

Diederik M. Roijers, Peter Vamplew, Shimon Whiteson, and Richard Dazeley. A survey of multi-objective sequential decision-making. *Journal of Artificial Intelligence Research*, 48(1):67–113, 2013.

Mark Rowland, Robert Dadashi, Saurabh Kumar, Remi Munos, Marc G. Bellemare, and Will Dabney. Statistics and samples in distributional reinforcement learning. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 5528–5536. PMLR, 2019.

Mark Rowland, Rémi Munos, Mohammad Gheshlaghi Azar, Yunhao Tang, George Ostrovski, Anna Harutyunyan, Karl Tuyls, Mark G. Bellemare, and Will Dabney. An analysis of quantile temporal-difference learning. *Journal of Machine Learning Research*, 25, 2024.

Gabor Szekely. E-statistics: The energy of statistical samples, Oct. 2002. Technical Report 2-16.

Thibaut Théate, Antoine Wehenkel, Adrien Bolland, Gilles Louppe, and Damien Ernst. Distributional reinforcement learning with unconstrained monotonic neural networks. *Neurocomputing*, 534(C):199–219, 2023.

C. Tsallis. Possible generalization of Boltzmann-Gibbs statistics. *Journal of Statistical Physics*, 52:479–487, 1988.

Harley Wiltzer, Jesse Farebrother, Arthur Gretton, and Mark Rowland. Foundations of multivariate distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 37, 2024.

Derek Yang, Li Zhao, Zichuan Lin, Tao Qin, Jiang Bian, and Tie-Yan Liu. Fully parameterized quantile function for distributional reinforcement learning. In *Advances in Neural Information Processing Systems*, volume 32, 2019.

Yaodong Yang, Jianye Hao, Mingyang Sun, Zan Wang, Changjie Fan, and Goran Strbac. Recurrent Deep Multiagent Q-Learning for Autonomous Brokers in Smart Grid. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*, volume 23, pp. 569–575, 7 2018.

Ruichong Zhang. Cramer Type Distances for Learning Gaussian Mixture Models by Gradient Descent, 2023. URL `https://arxiv.org/abs/2307.06753`.

Weijian Zhang, Jianshu Wang, and Yang Yu. Distributional Reinforcement Learning with Sample-set Bellman Update. In *IEEE, International Conference on Robotics and Automation (ICRA)*, 2024.

Fan Zhou, Chenfan Lu, Xiaocheng Tang, Fan Zhang, Zhiwei Qin, Jieping Ye, and Hongtu Zhu. Multi-objective distributional reinforcement learning for large-scale order dispatching. In *Proceedings of the IEEE International Conference on Data Mining (ICDM)*, pp. 1541–1546. IEEE, 2021.

A Representations of the random return

Random returns can be characterized in several ways. The most common representations are using the probability density function (PDF), the cumulative distribution function (CDF), the quantile function or the inverse CDF (QF) or the characteristic function (Φ). Specifically, denoting by $\mathcal{F}$ a latent functional space giving the neural networks architectures, the most used representations are the following.

CDF representation: $E = \{Z \mid \exists f \in \mathcal{F}, F_Z = f\}$. In that case, we have $F_{\gamma Z+r}(z) = F_Z\left(\frac{z-r}{\gamma}\right)$ and

$$\mathbb{E}[Z(x,a)] = \int_{-\infty}^{0} F_{Z(x,a)}(z)dz + \int_{0}^{+\infty} (1 - F_{Z(x,a)}(z))dz.$$

QF representation: $E = \{Z \mid \exists f \in \mathcal{F}, F_Z^{-1} = f\}$. In that case, we have $F_{\gamma Z+r}^{-1}(z) = \gamma F_Z^{-1}(z) + r$ and

$$\mathbb{E}[Z(x,a)] = \int_0^1 F_{Z(x,a)}^{-1}(t)dt.$$

PDF representation: $E = \{Z \mid \exists f \in \mathcal{F}, \eta_Z = f\}$. In that case, we have $\eta_{\gamma Z+r}(z) = \frac{1}{\gamma}\eta_Z\left(\frac{z-r}{\gamma}\right)$ and

$$\mathbb{E}[Z(x,a)] = \int_{\mathbb{R}} z\eta_{Z(x,a)}(z)dz.$$

Characteristic function representation:

$E = \{Z \mid \exists f \in \mathcal{F}, \Phi_Z = f\}$. We have $\Phi_{\gamma Z+r}(t) = e^{irt}\Phi_Z(\gamma t)$ and denoting by $\Im$ the imaginary part,

$$\mathbb{E}[Z(x,a)] = \Im\Phi'_{Z(x,a)}(0).$$

B Details on related work

Most methods (QR-DQN, IQN, FQF) use a quantile-based representation. The advantage of such an approach is that it is straightforward to derive the Monte-Carlo calculation of risk-aware policies with a distortion risk measure β using the functions largely studied in economics as follows :

$$Q(x,a) = \int_0^1 F^{-1}_{Z(x,a)}(\beta(\tau))d\tau.$$

B.1 Discrete representations

The principle of discrete algorithms is to take a mixture of atoms $\eta_Z = \sum_{k=1}^K w_k \delta_{\theta_k}$. Then there are two common points of view. Either we set the quantiles θ_k and we classify the weights using the KL divergence as this is done in C51, or we set the weights w_k (taking usually $w_k = \frac{1}{K}$) and we regress the quantiles (quantile regression) as this is done in QR-DQN using the loss :

$$\mathcal{L}_{QR}(Z_\theta, \tilde{Z}) = \sum_{k=1}^K \mathbb{E}[\rho_{\tau_k}(\tilde{Z} - \theta_k)],$$

where $\rho_\tau(u) = (\tau - 1_{u<0})u$ and $\tau_k = \sum_{i=1}^k w_i$.

Later, Rowland et al. (2019) apply the expectile regression in the same manner, by taking $\rho_\tau(u) = (\tau - 1_{u<0})u^2$ instead, leading to better practical results for their ER-DQN algorithm. This is not surprising, as the goal is to estimate the mean return and expectiles are generalizations of the mean, while quantiles are generalizations of the median.

Alternatively, Nguyen-Tang et al. (2020) use the MMD metric, instead of the previous losses, with the biased estimator

$$\mathcal{L}_{\text{MMD}}(X,Y) = \frac{1}{N^2}\sum_{i,j} k(X_i, X_j) + \frac{1}{M^2}\sum_{i,j} k(Y_i, Y_j) - \frac{2}{NM}\sum_{i,j} k(X_i, Y_j),$$

where $X = \{X_i\}_{i \in [N]}$ and $Y = \{Y_i\}_{i \in [M]}$ are sample sets generated from the distributions of Z_θ and $\tilde{Z}$. They use this estimator because it leads to less variance in practice, compensating the bias. However, they do not consider the possibility of exact computation in the case of parametric representations.

B.2 Continuous representations

Dabney et al. (2018a) use a fully-continuous representation of the distribution using neural networks which leads to a better approximation but they use quantile samples (Monte-Carlo) to estimate the Huber loss. Thus, this method adds unnecessary noise in the loss estimation depending on the choice of quantile fractions. This is why the state-of-the-art is now the FQF method that builds a neural network supervising the choice of the sampled quantile fractions.

C Advantages of parametric representations

We further specify, in the next sub-sections, some advantages of using parametric representations instead of the more common approaches.

C.1 Computational tractability

Gaussian mixtures (GM) provide a good trade-off between expressiveness and computational cost. Setting hyperparameters to the values in the original papers, we can assess, for various methods, the number of parameters as a function of the number of states $|\mathcal{A}|$. For our approach, denoted below GM-DQN, the number of mixture components is set to $K = 5$. Conformity is satisfied when the representation leads to a valid distribution. The comparison is reported in the following Table, which shows that GM provide the smallest number of parameters and preserve conformity.

Method comparison		
Method	Number of parameters	Conformity
GM-DQN	$15 \times \lvert\mathcal{A}\rvert$	✓
C51	$51 \times \lvert\mathcal{A}\rvert$	✓
QR-DQN	$200 \times \lvert\mathcal{A}\rvert$	✓
IQN	$576 \times \lvert\mathcal{A}\rvert$	×
FQF	$608 \times \lvert\mathcal{A}\rvert$	×

C.2 Tractability of risk measures

In DRL, we generally use risk-aware policies and variants to recover the agent policy from the Q-network using something like :

$$\pi(a \mid x) = \frac{1_{a\in\mathcal{K}(x)}}{|\mathcal{K}(x)|}, \quad \mathcal{K}(x) = \underset{a'\in\mathcal{A}}{\operatorname{argmax}} \mathcal{R}(Z(x,a')),$$

where $\mathcal{R}$ is a risk measure.

Let $\alpha \in (0,1)$ and X a random variable, the expectile $e_\alpha(X)$ is uniquely defined as the solution of (see Bellini & Bernardino (2017))

$$\alpha\mathbb{E}[(X-e_\alpha(X))_+] = (1-\alpha)\mathbb{E}[(X-e_\alpha(X))_-].$$

It is well-known that $e_{\frac{1}{2}}(X) = \mathbb{E}[X]$ and $\alpha \longmapsto e_\alpha(X)$ is increasing. It leads to an interesting greedy policy with $\mathcal{R} = e_\alpha$. This policy is risk-averse if $\alpha < \frac{1}{2}$ and risk-seeking otherwise. Similarly, we can take the quantile version $\mathcal{R} = q_\alpha$.

In our case with Gaussian mixtures, we can easily compute the expectiles and the quantiles using the Newton-Raphson method.

Theorem 8. *Let $X \sim \sum_{k=1}^K \pi_k\mathcal{N}(\mu_k,\sigma_k^2)$, we have $E_\alpha(e_\alpha(X)) = 0$ and $Q_\alpha(q_\alpha(X)) = 0$ where*

$$E_\alpha(x) := (1-2\alpha)\sum_{k=1}^K \pi_k\left(\sigma_k^2\mathcal{N}\left(x;\mu_k,\sigma_k^2\right) + (x-\mu_k)F_{\mathcal{N}(\mu_k,\sigma_k^2)}(x)\right) + \alpha\sum_{k=1}^K \pi_k(x-\mu_k),$$
$$Q_\alpha(x) := \sum_{k=1}^K \pi_k F_{\mathcal{N}(\mu_k,\sigma_k^2)}(x) - \alpha.$$

Moreover, we have

$$E'_\alpha(x) = (1-2\alpha)Q_\alpha(x) + 2\alpha(1-\alpha).$$

Another possible risk measure is $\mathcal{R}(X) = \sum_{k=1}^K \pi_k(\mu_k+\eta\sigma_k)$ for $X \sim \sum_{k=1}^K \pi_k\mathcal{N}(\mu_k,\sigma_k^2)$. All these proposed risk measures are popular in statistics and fully tractable using Gaussian mixtures.

As an illustration, Figure 4, shows the effect of a more (or less) risk-seeking policy on performance, still measured and shown as the average return. When considering the Assault ALE game, compared to the standard risk-neutral policy based on maximizing expectation, better returns are obtained with a more risk-averse policy (0.05 expectile), while worse returns are obtained with a more risk-seeking setting (0.95 expectile). This suggests that for this game, controlling actions in a more pessimistic manner is a better strategy. This should also be typically the case in life-dependent games. The choice of the expectile level impacts performance by influencing the exploration strategy but the effect of exploration also depends on the task, which may tolerate or not very risking options.

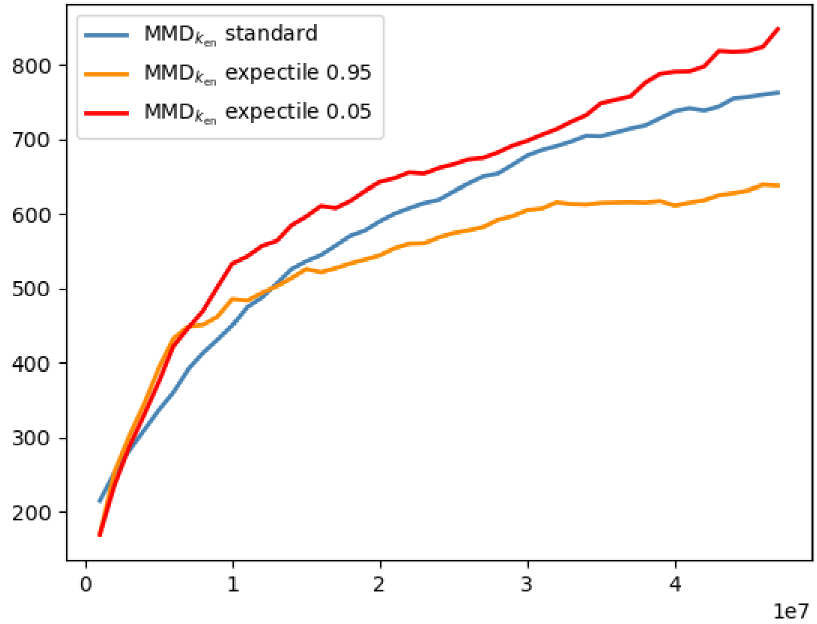


Figure 4: Gaussian mixture representations with the MMD energy distance. Effect of the risk-aware expectile policy on the ALE Assault game: Test return averaged over $500k$ testing frames for $47M$ training frames.

C.3 Multi-Objective Applications

Designing a reward function in RL can be challenging, as it must encapsulate and balance potentially conflicting objectives. In the classical RL setting, the agent observes a single scalar reward after interacting with the environment. This scalar reward implicitly reflects the weighted importance of different goals, guiding the agent to learn a policy π that maximizes it. However, once the policy is learned, it becomes difficult to disentangle and prioritize the individual contributions of each objective. This limitation occurs because the reward function does not explicitly represent the separate impact of each objective on the outcome.

Multi-objective RL (Roijers et al., 2013; Hayes et al., 2021) addresses this limitation by explicitly representing and ordering each objective i in a vector of cumulative returns, allowing the agent to learn a *single policy* that optimizes all objectives. Thus, action selection, guided by Pareto dominance, can be formalized as follows:

$$a^* = \arg\max_{a \in \mathcal{A}(x)} \mathbb{P}\left(\forall i, \forall a' \neq a \; \left(Z_i^\pi(x,a) \geq Z_i^\pi(x,a')\right)\right).$$

This approach enables the agent to select actions based on prioritized sub-goals without the need for retraining the policy each time one objective must be emphasized over the others. Recently, some works have explored parametric representations of returns (see *e.g.* Cai et al. (2023); Wiltzer et al. (2024)), opening up the possibility of representing multidimensional distributions suitable for multi-objective settings. GM are good candidates. Indeed, multivariate risk measures are well-known (Charpentier, 2018) and we have seen two types of probability metrics (Jensen-Tsallis and Mixture-Wasserstein) that are tractable with multivariate GM. This metrics could leverage the powerful ability of DRL to represent *aleatoric uncertainty*, allowing different strategies to handle it in the multi-objective RL setting.

D Proofs of main results

D.1 Sufficient conditions on D

Bellemare et al. (2017b) and Bellemare et al. (2023) introduced the following useful properties of D. D is said to be ideal if D satisfies the following to conditions,

$$\begin{aligned} &\textbf{(SI)} \quad \text{if } A \text{ is independent on } X \text{ and } Y \Longrightarrow D(A+X, A+Y) \leq D(X,Y) \\ &\textbf{(S)} \quad \text{considering } \lambda \in (0,1), \exists c > 0, \forall X, Y, \quad D(\lambda X, \lambda Y) \leq \lambda^c D(X,Y). \end{aligned}$$

If in addition D is p-convex, meaning (see Definition 4.24 in Bellemare et al. (2023)),

Definition 4. *Given $p > 1$, the probability metric D is p-convex if for any $\alpha \in (0,1)$ and distributions $\eta_1, \eta_2, \eta_1', \eta_2' \in \mathcal{P}(\mathbb{R})$, we have*

$$D^p(\alpha\eta_1 + (1-\alpha)\eta_2, \alpha\eta_1' + (1-\alpha)\eta_2') \leq \alpha D^p(\eta_1, \eta_1') + (1-\alpha) D^p(\eta_2, \eta_2') \ ,$$

then Theorem 4.25 of Bellemare et al. (2023) states that (SI) and (S) imply that, for any two return functions $\bar{\eta}_1$ and $\bar{\eta}_2$,

$$\overline{D}(\mathcal{T}^\pi \bar{\eta}_1, \mathcal{T}^\pi \bar{\eta}_2) \leq \gamma^c \overline{D}(\bar{\eta}_1, \bar{\eta}_2).$$

Note that the assumptions in Theorem 4.25 of Bellemare et al. (2023) use the fact that D is regular, which is equivalent to satisfying (SI) (see Definition 4.23 therein), and that D is c-homogeneous which is (S) but where the inequality is replaced by an equality. However, it is easy to see from the proof of Theorem 4.25, that the inequality given by the (S) condition is enough.

D.2 Proof of Theorem 1

Proof. The proof is straightforward combining Theorems 4 and 6 that show respectively that $\mathcal{T}^\pi$ is a contraction and that the projection is non-expansive. So that for two return function $\bar{\eta}_1$ and $\bar{\eta}_2$,

$$\overline{MW}_2(\overline{\Pi^{w*}_{MW2}}\mathcal{T}^\pi \bar{\eta}_1, \overline{\Pi^{w*}_{MW2}}\mathcal{T}^\pi \bar{\eta}_2) \leq \overline{MW2}(\mathcal{T}^\pi \bar{\eta}_1, \mathcal{T}^\pi \bar{\eta}_2) \leq \gamma \overline{MW}_2(\bar{\eta}_1, \bar{\eta}_2).$$

□

D.3 Proof and specifications on Theorem 2

Proof. We assume $\mathbb{E}[\omega(A)] < \infty$, although inequality (1) can still makes sense if the right-hand site of (1) is infinite. With η_A the PDF of A and A independent on X and Y, the PDF of $X + A$ (resp. $Y + A$) is the convolution $\eta_A * \eta_X$ (resp. $\eta_A * \eta_Y$). This latter condition is also important as it prevents to use inequality (1) with X set to $X + A$, Y set to $Y + A$ and A to $-A$, which would lead to the opposite inequality and then to an equality in (1). We thus obtain, $JT_{\omega,2}(A+X, A+Y) = \|(\eta_X - \eta_Y) * \eta_A\|^2_{L^2(\omega)}$. Hence using the Young inequality (Bogachev, 2007), we get (1) as follows,

$$JT_{\omega,2}(A+X, A+Y) \leq \|\eta_X - \eta_Y\|^2_{L^2(\omega)} \|\eta_A\|^2_{L^1(\omega)} \leq JT_{\omega,2}(X,Y)\mathbb{E}[\omega(A)]^2.$$

Then, using $\eta_{\lambda X}(x) = \lambda^{-1}\eta_X(\lambda^{-1}x)$, we obtain

$$JT_{\omega,2}(\lambda X, \lambda Y) = \lambda^{-1} \int_{\mathbb{R}} \omega(\lambda x) |\eta_X(x) - \eta_Y(x)|^2 dx \tag{5}$$

from which (2) follows. It is straightforward to see that $\mathrm{JT}_{1,2}$ does not satisfy (S) according to (5). Indeed, applying (5) with $\omega = 1$, leads to $JT_{1,2}(\lambda X, \lambda Y) = \lambda^{-1} JT_{1,2}(X,Y)$. So that if (S) was satisfied for $\mathrm{JT}_{1,2}$, this would mean $\lambda^{c+1} \geq 1$ for some $c > 0$ and $\lambda \geq 1$, which conflicts with the assumption $\lambda \in (0,1)$. □

Furthermore, when considering $\omega(x) = x^2$, inequality (2) in the paper leads to

$$JT_{x^2,2}(\lambda X, \lambda Y) \leq \lambda JT_{x^2,2}(X,Y) \ .$$

For $\lambda \in (0,1)$, it follows that condition (S) is satisfied with $c = 1$. In contrast, (1) implies that

$$JT_{x^2,2}(A+X, A+Y) \leq JT_{x^2,2}(X,Y)\mathbb{E}[A^2]^2 \; .$$

If $\mathbb{E}[A^2] > 1$, this may prevent (SI) to be satisfied since the Young inequality is optimal.

For instance, if A is constant, $A = a$, with $a > \max\left(0, -\frac{2JT_{x,2}(X,Y)}{JT_{1,2}(X,Y)}\right)$, it comes

$$JT_{x^2,2}(A+X, A+Y) = \int_{\mathbb{R}} x^2(\eta_X(x-a) - \eta_Y(x-a))^2 dx = \int_{\mathbb{R}} (x+a)^2(\eta_X(x) - \eta_Y(x))^2 dx,$$

which leads to

$$JT_{x^2,2}(A+X, A+Y) = a(aJT_{1,2}(X,Y) + 2JT_{x,2}(X,Y)) + JT_{x^2,2}(X,Y) > JT_{x^2,2}(X,Y).$$

D.4 Proof of Theorem 3

Let us first note that the result is easy to check if $\mathbb{V}[A] = 0$. In that case A is a constant random variable and for any scalar a, we can show that since $\eta_{a+\gamma X}(x) = \eta_X(\frac{x-a}{\gamma})$, then for $\gamma \in (0,1)$,

$$JT_{1,2}(a+\gamma X, a+\gamma Y) = JT_{1,2}(\gamma X, \gamma Y) = \gamma^{-1} JT_{1,2}(X,Y) > JT_{1,2}(X,Y).$$

More generally using the mean/variance decomposition $A = \mathbb{E}[A] + \mathbb{V}[A]^{1/2}U$, where U is a standardized random variable with $\mathbb{E}[U] = 0$ and $\mathbb{V}[U] = 1$, it follows that

$$JT_{1,2}(A+\gamma X, A+\gamma Y) = JT_{1,2}(\mathbb{V}[A]^{1/2}U + \gamma X, \mathbb{V}[A]^{1/2}U + \gamma Y),$$

where the right-hand side is independent on the mean of A.

Then, if $\mathbb{V}[A] \to 0$, the distribution $p_{\mathbb{V}[A]^{1/2}U} \to \delta_0$, where δ_0 indicates the Dirac mass in 0. Using that the convolution operator is a bilinear continuous operator (owing to the weak Young inequality) and that $\delta_0 * p = p * \delta_0 = p$, we can deduce that when $\mathbb{V}[A] \to 0$ then

$$\begin{aligned} JT_{1,2}(\mathbb{V}[A]^{1/2}U + \gamma X, \mathbb{V}[A]^{1/2}U + \gamma Y) =& ||\eta_{\mathbb{V}[A]^{1/2}U} * (\eta_{\gamma X} - \eta_{\gamma Y})||_2^2 \\ &\to ||(\eta_{\gamma X} - \eta_{\gamma Y})||_2^2 = JT_{1,2}(\gamma X, \gamma Y) \; . \end{aligned}$$

More formally, let us decompose the difference

$$\begin{aligned} JT_{1,2}(A+\gamma X, A+\gamma Y) - JT_{1,2}(X,Y) =& JT_{1,2}(A+\gamma X, A+\gamma Y) - JT_{1,2}(\gamma X, \gamma Y) \\ &+ JT_{1,2}(\gamma X, \gamma Y) - JT_{1,2}(X,Y) \; . \end{aligned}$$

The second difference is $JT_{1,2}(\gamma X, \gamma Y) - JT_{1,2}(X,Y) = (\frac{1}{\gamma} - 1)JT_{1,2}(X,Y)$, which is strictly positive if $\gamma \in (0,1)$ and $JT_{1,2}(X,Y) \neq 0$ or equivalently $\eta_X \neq \eta_Y$. For the first term, using continuity,

$$\forall \epsilon > 0, \exists \epsilon' > 0, \text{ so that } \mathbb{V}[A] \leq \epsilon' \Rightarrow |JT_{1,2}(A+\gamma X, A+\gamma Y) - JT_{1,2}(\gamma X, \gamma Y)| \leq \epsilon \; .$$

Choosing $\epsilon < (\frac{1}{\gamma} - 1)JT_{1,2}(X,Y)$ and $\sigma_{max} = \epsilon'$ ends the proof.

D.5 Proof of Lemma 1

Let A be independent of X and Y and all these random variables distributed according to a mixture of distributions in $\mathcal{L}$, denoted respectively by $\eta_Z = \sum_{k=1}^{K^Z} \pi_k^Z \ell_k^Z$, where Z represents in turn X, Y or A and $\ell_k^Z \in \mathcal{L}$. For $Z = X$ and $Z = Y$, by distributivity of convolution and summing stability of $\mathcal{L}$, $Z + A$ also follows a mixture of elements in $\mathcal{L}$ given by,

$$\eta_{Z+A} = \eta_Z * \eta_A = \sum_{k=1}^{K^Z} \sum_{i=1}^{K^A} \pi_k^Z \pi_i^A \; (\ell_k^Z * \ell_i^A) = \sum_{k=1}^{K^Z} \sum_{i=1}^{K^A} \pi_{ki}^{Z+A} \ell_{ki}^{Z+A}.$$

Let $w = (w_{kl}, k \in [K^X], l \in [K^Y])$ be a discrete distribution. For $w \in \Pi(\pi^X, \pi^Y)$, the marginals of w are π^X and π^Y. For such a w, define the discrete distribution $\tilde{w} = (w_{klij}, k \in [K^X], l \in [K^Y], i, j \in [K^A])$ where,

$$\begin{aligned} w_{klij} &= w_{kl}\ \pi_i^A \quad \text{ if } i = j \\ &= 0 \quad \text{otherwise} \end{aligned}$$

The set of such distributions is denoted by

$$\Pi' = \{\tilde{w},\ \text{s.t } w \in \Pi(\pi^X, \pi^Y)\}.$$

For an element in Π', the marginals are respectively π^{X+A} and π^{Y+A} since,

$$\sum_{l=1}^{K^Y} \sum_{j=1}^{K^A} w_{kl}\ \pi_i^A\ \delta_{\{i=j\}} = \pi_i^A \sum_{l=1}^{K^Y} w_{kl} = \pi_i^A \pi_k^X = \pi_{ki}^{X+A}$$

and similarly,

$$\sum_{k=1}^{K^X} \sum_{i=1}^{K^A} w_{kl}\ \pi_i^A\ \delta_{\{i=j\}} = \pi_j^A \sum_{k=1}^{K^X} w_{kl} = \pi_j^A \pi_l^Y = \pi_{lj}^{Y+A}\ .$$

It follows that $\Pi' \subset \Pi(\pi^{X+A}, \pi^{Y+A})$.

Let us first prove the (SI) property. By definition of $\mathrm{MW}_{D,\mathcal{L}}^p$,

$$\begin{aligned} \mathrm{MW}_{D,\mathcal{L}}^p(X + A, Y + A) &= \mathrm{MW}_{D,\mathcal{L}}^p(\eta_{X+A}, \eta_{Y+A}) \\ &= \min_{\tilde{w} \in \Pi(\pi^{X+A}, \pi^{Y+A})} \sum_{k,l,i,j} w_{klij}\ D^p(\ell_{ki}^{X+A}, \ell_{lj}^{Y+A}) \\ &\leq \min_{\tilde{w} \in \Pi'} \sum_{k,l} w_{kl} \left(\sum_i \pi_i^A\ D^p(\ell_{ki}^{X+A}, \ell_{li}^{Y+A}) \right) \end{aligned}$$

Using the (SI) property of D, we have $D(\ell_{ki}^{X+A}, \ell_{li}^{Y+A}) \leq D(\ell_k^X, \ell_l^Y)$ from which we deduce the (SI) property for $\mathrm{MW}_{D,\mathcal{L}}^p$, that is,

$$\mathrm{MW}_{D,\mathcal{L}}^p(X + A, Y + A) \leq \min_{\tilde{w} \in \Pi'} \sum_{k,l} w_{kl} D^p(\ell_k^X, \ell_l^Y)) = \mathrm{MW}_{D,\mathcal{L}}^p(X, Y)\ .$$

For the (S) property, for all $\lambda > 0$, we have $\eta_{\lambda Z}(x) = \sum_{k=1}^{K^Z} \pi_k^Z\ \ell_{k,\lambda}^Z$ with $\ell_{k,\lambda}^Z = \ell_k^{\lambda Z}$, since $\mathcal{L}$ is stable by scaling. Thus, using the (S) property of D, we obtain

$$\begin{aligned} \mathrm{MW}_{D,\mathcal{L}}^p(\lambda X, \lambda Y) &= \min_{w \in \Pi(\pi^X, \pi^Y)} \sum_{k,l} w_{kl}\ D^p(\ell_k^{\lambda X}, \ell_l^{\lambda Y}) \\ &\leq \min_{w \in \Pi(\pi^X, \pi^Y)} \sum_{k,l} w_{kl}\ \lambda^{cp}\ D^p(\ell_k^X, \ell_l^Y) = \lambda^{cp}\ \mathrm{MW}_{D,\mathcal{L}}^p(X, Y)\ . \end{aligned}$$

This achieves the proof that $\mathrm{MW}_{D,\mathcal{L}}^p$ is ideal.

D.6 Proof of Theorem 4

As mentioned in Section D.1, Theorem 4.25 of Bellemare et al. (2023) can be used to get the desired contraction result. Note that although Theorem 4.25 of Bellemare et al. (2023) is stated for probability metrics, it is easy to check that the proof does not use any particular property of probability metrics so it is still valid for quasi-metrics or discrepancies.

Thus, as Lemma 1 above implies the (S) and (SI) conditions, it remains to show that $\mathrm{MW}_{D,\mathcal{L}}^p$ is 1-convex. This follows from the convexity of the discrete Wasserstein distance, which we prove below for completeness.

Let $\alpha \in (0,1)$. Let for $i=1,2$, $\eta_i = \sum_{i=1}^{K_i} \pi_{ik}\ell_{ik}$ and $\eta'_i = \sum_{i=1}^{K'_i} \pi'_{ik}\ell'_{ik}$ be four mixtures distributions with elements in $\mathcal{L}$. By construction, distributions of the form $\alpha\ \eta_i + (1-\alpha)\ \eta'_i$ are also mixtures of elements in $\mathcal{L}$. More specifically,

$$\alpha\ \eta_i + (1-\alpha)\ \eta'_i = \sum_{k=1}^{K_i+K'_i} \alpha\ \pi_{ik}\ \delta_{\{k\leq K_i\}}\ell_{ik} + (1-\alpha)\ \pi'_{ik}\ \delta_{\{k>K_i\}}\ell'_{ik} = \sum_{k=1}^{K_i+K'_i} \hat{\pi}_{ik}\ \hat{\ell}_{ik}\ .$$

Moreover, defining

$$\hat{\Pi} = \{(\alpha\ \delta_{\{k\leq K_1, l\leq K_2\}} w_{kl} + (1-\alpha)\ \delta_{\{k>K_1, l>K_2\}} w'_{kl})_{kl} \text{ with } w \in \Pi(\pi_1,\pi_2) \text{ and } w' \in \Pi(\pi'_1,\pi'_2)\},$$

we have $\hat{\Pi} \subset \Pi(\hat{\pi}_1,\hat{\pi}_2)$. It follows,

$$\begin{aligned}
\mathrm{MW}^p_{D,\mathcal{L}}(\alpha\ \eta_1 + (1-\alpha)\ \eta'_1, \alpha\ \eta_2 + (1-\alpha)\ \eta'_2) &= \min_{w\in\Pi(\hat{\pi}_1,\hat{\pi}_2)} \sum_{k=1}^{K_1+K'_1}\sum_{l=1}^{K_2+K'_2} w_{kl} D(\hat{\ell}_{1k},\hat{\ell}_{2l})^p \\
&\leq \min_{w\in\hat{\Pi}} \sum_{k=1}^{K_1+K'_1}\sum_{l=1}^{K_2+K'_2} w_{kl} D(\hat{\ell}_{1k},\hat{\ell}_{2l})^p \\
&\leq \min_{w\in\Pi(\pi_1,\pi_2)} \sum_{k=1}^{K_1}\sum_{l=1}^{K_2} \alpha w_{k,l} D(\ell_{1k},\ell_{2l})^p \\
&\quad + \min_{w'\in\Pi(\pi'_1,\pi'_2)} \sum_{k=1}^{K'_1}\sum_{l=1}^{K'_2} (1-\alpha) w'_{k,l} D(\ell'_{1k},\ell'_{2l})^p \\
&\leq \alpha \mathrm{MW}^p_{D,\mathcal{L}}(\eta_1,\eta_2) + (1-\alpha)\mathrm{MW}^p_{D,\mathcal{L}}(\eta'_1,\eta'_2).
\end{aligned}$$

D.7 Proof of Theorem 5

The goal is to investigate the reliability of performing a stochastic gradient descent with respect to θ over a MW_2 loss. In practice, we rather equivalently consider the square of the metric, MW_2^2 to avoid fractional exponents. We thus consider the estimate $\nabla_\theta MW_2^2(\hat{\eta}_M,\eta_\theta)$ of the gradient $\nabla_\theta MW_2^2(\eta_X,\eta_\theta)$ and show that is in general biased. Write $\bar{X} = \{X_m\}_{m\in[M]}$. Recall that $\Pi(\pi,\pi')$ denotes the set of discrete joint distributions $w = (w_{k\ell}, k\in[K], \ell\in[K'])$, whose marginals are $\pi = (\pi_1,\ldots,\pi_K)$ and $\pi' = (\pi'_1,\ldots,\pi'_{K'})$,

$$\Pi(\pi,\pi') = \{w = (w_{k\ell})_{k,\ell}, s.t. \sum_{k,\ell} w_{k\ell} = 1, \pi_k = \sum_\ell w_{k\ell}, \pi'_\ell = \sum_k w_{k\ell}\}\ .$$

Using Definition 4 of the squared MW_2 metric, we denote by $w^*(\bar{X}) = (w^*_{km}, k\in[K], m\in[M])$ the optimal coupling defining $\mathrm{MW}_2^2(\hat{\eta}_M,\eta_\theta)$ and similarly by $\tilde{w}^* = (\tilde{w}^*_{k\ell}, k\in[K], \ell\in[K_X])$ the optimal coupling defining $\mathrm{MW}_2^2(\eta_X,\eta_\theta)$.

We can then write,

$$\begin{aligned}
MW_2^2(\hat{\eta}_M,\eta_\theta) &= \sum_{k,m} w^*_{km}(\bar{X})\left((X_m-\mu_{k,\theta})^2 + \sigma^2_{k,\theta}\right) && (6)\\
&= \sum_k \pi_{k,\theta}(\mu^2_{k,\theta}+\sigma^2_{k,\theta}) + \sum_m X_m^2/M - 2\sum_{k,m} w^*_{km}(\bar{X})X_m\mu_{k,\theta} && (7)\\
& && (8)
\end{aligned}$$

Then,

$$\begin{aligned}
\mathbb{E}\left[\nabla_\theta MW_2^2(\hat{\eta}_M,\eta_\theta)\right] &= \nabla_\theta\left(\sum_k \pi_{k,\theta}(\mu^2_{k,\theta}+\sigma^2_{k,\theta})\right) - 2\nabla_\theta\left(\sum_{k,m}\mathbb{E}[w^*_{km}(\bar{X})X_m]\ \mu_{k,\theta}\right) && (9)\\
&= \nabla_\theta\left(\sum_k \pi_{k,\theta}(\mu^2_{k,\theta}+\sigma^2_{k,\theta})\right) - 2\nabla_\theta\left(\sum_k \mu_{k,\theta}\mathbb{E}[\sum_m w^*_{km}(\bar{X})X_m]\right) && (10)
\end{aligned}$$

While the target distance is

$$MW_2^2(\eta_X, \eta_\theta) = \sum_{k,\ell} \tilde{w}^*_{k\ell} \left((\mu_{\ell,X} - \mu_{k,\theta})^2 + (\sigma_{\ell,X} - \sigma_{k,\theta})^2 \right) \tag{11}$$

from which we can derive that

$$\nabla_\theta MW_2^2(\eta_X, \eta_\theta) = \nabla_\theta \left(\sum_k \pi_{k,\theta} (\mu_{k,\theta}^2 + \sigma_{k,\theta}^2) \right) - 2\nabla_\theta \left(\sum_{k,\ell} \tilde{w}^*_{k\ell} (\mu_{\ell,X}\mu_{k,\theta} + \sigma_{\ell,X}\sigma_{k,\theta}) \right) . \tag{12}$$

It follows that the difference $\mathbb{E}\left[\nabla_\theta MW_2^2(\hat{\eta}_M, \eta_\theta)\right] - \nabla_\theta MW_2^2(\eta_X, \eta_\theta)$ is equal to

$$2\nabla_\theta \left(\sum_{k,\ell} \tilde{w}^*_{k\ell} (\mu_{\ell,X}\mu_{k,\theta} + \sigma_{\ell,X}\sigma_{k,\theta}) \right) - 2\nabla_\theta \left(\sum_k \mu_{k,\theta} \mathbb{E}[\sum_m w^*_{km}(\bar{X}) X_m] \right) \tag{13}$$

which is in general non zero. Standard stochastic gradient optimization does not come with standard guarantees when applied to a MW_2^2 loss. However, several solutions using biased gradients have been investigated and may be possible without too restrictive assumptions; see the recent review of Demidovich et al. (2023).

D.8 Proof of Theorem 6

Let us consider the projection on $\mathcal{M}_K$ the space of GM with a fixed number K of components. The proof consists mainly is exhibiting the subset W_K^*. Recall that $\mathcal{M}$ denotes the set of all GM. The goal is to show that when considering the MW_2^2 metric, the projection on $\mathcal{M}_K$ is non expansive. This projection is defined for any GM distribution $\eta \in \mathcal{M}$ as $\arg\min_{\eta' \in \mathcal{M}_K} MW_2^2(\eta, \eta')$. As the univariate MW_2^2 metric is defined with the Euclidean distance as cost function, we can first make use of standard results to explicit this projection.

Consider for some arbitrary $L \in \mathbb{N}$, $\eta = \sum_{\ell=1}^L \pi_\ell \mathcal{N}(\mu_\ell, \sigma_\ell^2)$ and $\eta' = \sum_{k=1}^K \pi'_k \mathcal{N}(\mu'_k, \sigma_k'^2)$ two univariate mixtures in $\mathcal{M}_L$ and $\mathcal{M}_K$ respectively. Denote $\pi = (\pi_1, \dots, \pi_L)$, $\pi' = (\pi'_1, \dots, \pi'_K)$ and the set of couplings between π and π' by

$$\Pi(\pi, \pi') = \{w = (w_{\ell k})_{\ell,k}, s.t. \sum_{\ell,k} w_{\ell k} = 1, \pi_\ell = \sum_k w_{\ell k}, \pi'_k = \sum_\ell w_{\ell k}\}.$$

Recall that, for any random variables X and Y, the characterization of $\mathbb{E}[X \mid Y]$ as the unique L^2 projection of X leads to

$$\mathbb{E}[(X - Y)^2] \geq \mathbb{E}[(X - \mathbb{E}[X \mid Y])^2] \; . \tag{14}$$

For $w = (w_{\ell k})_{\ell,k} \in \Pi(\pi, \pi')$, we then consider two variables X and Y so that $(X, Y) \sim \hat{\eta}_{LK}$ where $\hat{\eta}_{LK} = \sum_{\ell,k} w_{\ell k} \delta_{(x_\ell, y_k)}$. For all $\{x_\ell\}_\ell, \{y_k\}_k, w \in \Pi(\pi, \pi')$, inequality (14) writes,

$$\sum_{\ell,k} w_{\ell k} (x_\ell - y_k)^2 \geq \sum_{\ell,k} w_{\ell k} \left(x_\ell - \sum_{\ell'} \frac{w_{\ell' k}}{\pi'_k} x_{\ell'} \right)^2 , \tag{15}$$

with equality if and only if $y_k = \sum_{\ell'} \frac{w_{\ell' k}}{\pi'_k} x_{\ell'}$ for all k.

Then using that $\mathrm{MW}_2^2(\eta, \eta')$ is defined as the minimum over $w \in \Pi(\pi, \pi')$ of

$$\Lambda(w, \eta, \eta') = \sum_{\ell,k} w_{\ell k} \left((\mu_\ell - \mu'_k)^2 + (\sigma_\ell - \sigma'_k)^2 \right) ,$$

we can replace x and y in (15) with successively the means and then the standard deviations of the two mixtures components to get,

$$\Lambda(w,\eta,\eta') \geq \sum_{\ell,k} w_{\ell k}\left(\left(\mu_\ell - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\mu_{\ell'}\right)^2 + \left(\sigma_\ell - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\sigma_{\ell'}\right)^2\right).$$

It follows that,

$$MW_2^2(\eta,\eta') \geq \min_{w\in\Pi(\pi,\pi')}\sum_{\ell,k} w_{\ell k}\left(\left(\mu_\ell - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\mu_{\ell'}\right)^2 + \left(\sigma_\ell - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\sigma_{\ell'}\right)^2\right) ,$$

and then that,

$$\min_{\eta'\in\mathcal{M}_K} MW_2^2(\eta,\eta') \geq \min_{\pi'}\min_{w\in\Pi(\pi,\pi')}\sum_{\ell,k} w_{\ell k}\left(\left(\mu_\ell - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\mu_{\ell'}\right)^2 + \left(\sigma_\ell - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\sigma_{\ell'}\right)^2\right) ,$$

Then denoting by $w^*(\eta)$ the coupling where the minimum above is reached, we define $\tilde{\mu}_k(\eta) = \sum_{\ell'}\frac{w^*_{\ell' k}(\eta)}{\tilde{\pi}_k(\eta)}\mu_{\ell'}$, $\tilde{\sigma}_k(\eta) = \sum_{\ell'}\frac{w^*_{\ell' k}(\eta)}{\tilde{\pi}_k(\eta)}\sigma_{\ell'}$, $\tilde{\pi}_k(\eta) = \sum_{\ell'} w^*_{\ell' k}(\eta)$ and consider the mixture $\tilde{\eta}_{w^*(\eta)} \in \mathcal{M}_K$, $\tilde{\eta}_{w^*(\eta)} = \sum_k \tilde{\pi}_k(\eta)\mathcal{N}(\tilde{\mu}_k(\eta), \tilde{\sigma}_k^2(\eta))$. Thus we obtain the projection operator $\Pi^{w^*}_{MW_2}\eta = \tilde{\eta}_{w^*(\eta)}$ that satisfies

$$MW_2^2(\eta, \Pi^{w^*}_{MW_2}\eta) = \min_{\eta'\in\mathcal{M}_K} MW_2^2(\eta,\eta').$$

In addition, $\tilde{\eta}_{w^*(\eta)}$ is a mixture whose expectation is the same as that of η, *i.e.* $\sum_\ell \pi_\ell\mu_\ell$ so that $\mathbb{E}[\Pi^{w^*}Z(x,a)] = \mathbb{E}[Z(x,a)]$ if the PDF of $Z(x,a)$ (resp. $\Pi^{w^*}Z(x,a)$) is η (resp. $\Pi^{w^*}\eta$) so the projection is mean-preserving.

Similarly, it is easy to show the non-expansion. For two mixtures η and η' and $w \in \Pi(\pi,\pi')$, we consider a random vector $(X_0, X_1, Y_0, Y_1) \sim \sum_{i,j,k,\ell}\tilde{w}_{i,j,k,\ell}\delta_{(x_i, x'_j, y_k, y'_\ell)}$ such that

$$\begin{aligned}
\forall i,j,\quad & \sum_{k,\ell}\tilde{w}_{i,j,k,\ell} = w^*_{ij}(\eta),\\
\forall k,\ell,\quad & \sum_{i,j}\tilde{w}_{i,j,k,\ell} = w^*_{k\ell}(\eta'),\\
\forall i,k,\quad & \sum_{j,\ell}\tilde{w}_{i,j,k,\ell} = w_{ik},
\end{aligned}$$

with X_0 (resp. Y_0) independent of Y_1 (resp. X_1). Using Cauchy-Schwarz inequality, we obtain

$$\mathbb{E}[X_0 - Y_0 \mid (X_1, Y_1) = (x'_j, y'_\ell)]^2 \leq \mathbb{E}[(X_0 - Y_0)^2 \mid (X_1, Y_1) = (x'_j, y'_\ell)].$$

Moreover, since we have independence, we get

$$\mathbb{E}[X_0 - Y_0 \mid (X_1, Y_1) = (x'_j, y'_\ell)]^2 = \mathbb{E}[X_0 \mid X_1 = x'_j] - \mathbb{E}[Y_0 \mid Y_1 = y'_\ell],$$

so the following inequality holds,

$$\forall j,\ell,\ \left(\sum_i \frac{w^*_{ij}(\eta)}{\tilde{\pi}_j(\eta)}x_i - \sum_k \frac{w^*_{k\ell}(\eta')}{\tilde{\pi}_\ell(\eta')}y_k\right)^2 \leq \sum_{i,k}\mathbb{P}((X_0,Y_0) = (x_i,y_k) \mid (X_1,Y_1) = (x'_j, y'_\ell))(x_i - y_k)^2.$$

Hence we obtain

$$\sum_{j,\ell}\mathbb{P}((X_1,Y_1) = (x'_j,y'_\ell))\left(\sum_i \frac{w^*_{ij}(\eta)}{\tilde{\pi}_j(\eta)}x_i - \sum_k \frac{w^*_{k\ell}(\eta')}{\tilde{\pi}_\ell(\eta')}y_k\right)^2 \leq \sum_{i,k} w_{ik}(x_i - y_k)^2, \tag{16}$$

for all x, y, $w \in \Pi(\pi, \pi')$. Since $(\mathbb{P}((X_1, Y_1) = (x'_j, y'_\ell)))_{j,\ell} \in \Pi(\tilde{\pi}(\eta), \tilde{\pi}(\eta'))$, we can replace the x_i's' and y_i's' in (16) by the means and standard deviations of the two mixtures components to get

$$MW_2^2(\Pi_{MW_2}^{w^*}\eta, \Pi_{MW_2}^{w^*}\eta') \leq MW_2^2(\eta, \eta'),$$

thus $\Pi_{MW_2}^{w^*}$ is a non-expansion with respect to MW_2.

D.9 Unbiasedness property

If the sample estimates of the projected distributional Bellman operator are unbiased, then it is easy to conclude to the convergence of the associated TD algorithm. This is how the convergence of the categorical TD algorithm is typically shown (see Bellemare et al. (2023)). However, this is not possible in our case.

Indeed, define for all return function $\bar{\eta}$ and for all $x \in \mathcal{X}$,

$$\begin{aligned}
\eta^\pi(x) &= \sum_{a \in \mathcal{A}} \bar{\eta}(x, a)\pi(a \mid x), \\
\rho^\pi(x) &= \sum_{a \in \mathcal{A}} \rho(x, a)\pi(a \mid x), \\
P^\pi(\cdot \mid x) &= \sum_{a \in \mathcal{A}} P(\cdot \mid x, a)\pi(a \mid x).
\end{aligned}$$

The standard property of unbiasedness defined in p. 166 of Bellemare et al. (2023), which would amount to check the following equality, with our projection $\Pi_{\mathrm{MW}_2}^{w^*}$ defined previously,

$$\forall x \in \mathcal{X}, \quad \int_{\mathcal{X}} \int_{\mathbb{R}} \Pi_{\mathrm{MW}_2}^{w^*}((r + \gamma Id)_{\#}\eta^\pi(x'))\rho^\pi(x)(r)P^\pi(x' \mid x)drdx' = \Pi_{\mathrm{MW}_2}^{w^*}\mathcal{T}^\pi\eta^\pi(x),$$

is not satisfied, as it is also not the case for quantile temporal difference (see example 6.3 of Bellemare et al. (2023)).

In fact, we can show instead that

$$\int_{\mathcal{X}} \int_{\mathbb{R}} \Pi_{\mathrm{MW}_2}^{w^*}((r + \gamma Id)_{\#}\eta^\pi(x'))\rho^\pi(x)(r)P^\pi(x' \mid x)drdx' = \mathcal{T}^\pi\Pi_{\mathrm{MW}_2}^{w^*}\eta^\pi(x). \tag{17}$$

Thus we indeed have a bias since the result $\mathcal{T}^\pi\Pi^{w^*}\eta^\pi$ differs from $\Pi^{w^*}\mathcal{T}^\pi\eta^\pi$.

Proof of (17). To compute $\Pi^{w^*}((r + \gamma Id)_{\#}\eta^\pi(x'))$, we use that $w^*((r + \gamma Id)_{\#}\eta)$ is defined as

$$w^*((r + \gamma Id)_{\#}\eta) \in \arg \min_{w, \sum_k w_{\ell k} = \pi_\ell} \sum_{\ell,k} w_{\ell k}\left(\left(\gamma\mu_\ell + r - \sum_{\ell'} \frac{w_{\ell' k}}{\sum_i w_{ik}}(\gamma\mu_{\ell'} + r)\right)^2 + \left(\gamma\sigma_\ell - \sum_{\ell'} \frac{w_{\ell' k}}{\sum_i w_{ik}}(\gamma\sigma_{\ell'})\right)^2\right).$$

It follows that $w^*((r + \gamma Id)_{\#}\eta)$ can be chosen the same as $w^*(\eta)$ and thus that $\Pi^{w^*}((r + \gamma Id)_{\#}\eta^\pi(x')) = (r + \gamma Id)_{\#}\Pi^{w^*}(\eta^\pi(x'))$, which leads to (17) as desired.

D.10 Generalization to the multidimensional case

D.10.1 General metric properties

We define the F^2 probability metric between two multivariate Gaussian distributions in dimension d, $g_1 = \mathcal{N}_d(\mu_1, \Sigma_1)$ and $g_2 = \mathcal{N}_d(\mu_2, \Sigma_2)$ as

$$F^2(g_1, g_2) = \|\mu_1 - \mu_2\|^2 + Tr\left(\left(\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}}\right)^2\right).$$

This semi-metric is an upper-bound of the 2-Wasserstein distance making it an appealing alternative with a projection easier to manipulate. Indeed, we have

$$Tr((\Sigma_1^{\frac{1}{2}} - \Sigma_2^{\frac{1}{2}})^2) = Tr(\Sigma_1 + \Sigma_2 - 2\Sigma_1^{\frac{1}{2}}\Sigma_2^{\frac{1}{2}}) \geq Tr(\Sigma_1 + \Sigma_2 - 2(\Sigma_1^{\frac{1}{2}}\Sigma_2\Sigma_1^{\frac{1}{2}})^{\frac{1}{2}}),$$

using the classical inequality $Tr((AA^T)^{\frac{1}{2}}) \geq Tr(A)$ with $A = \Sigma_1^{\frac{1}{2}}\Sigma_2^{\frac{1}{2}}$. Moreover, it is ideal according to the following theorem 9 so we can apply Theorem 4 to get that $\mathcal{T}^\pi$ is a contraction mapping with respect to $\mathrm{MW}^2_{F^2,\mathcal{M}^d}$ if the reward is a multivariate Gaussian mixture for each state-action pair.

Theorem 9. *F is ideal.*

Proof. It is clear that F satisfies **(S)**. In addition, for all $i \in \{1,2\}$ and $g' = \mathcal{N}_d(\mu', \Sigma')$, we have $g' * g_i = \mathcal{N}_d(\mu_i + \mu', \Sigma_i + \Sigma')$ so we get

$$F^2(g' * g_1, g' * g_2) - F^2(g_1, g_2) = 2Tr\left(\Sigma' + \Sigma_1^{\frac{1}{2}}\Sigma_2^{\frac{1}{2}} - (\Sigma_1 + \Sigma')^{\frac{1}{2}}(\Sigma_2 + \Sigma')^{\frac{1}{2}}\right). \tag{18}$$

Moreover, The function $\phi : (M, N, K) \longmapsto Tr(M^{\frac{1}{4}}K^TN^{\frac{1}{2}}KM^{\frac{1}{4}})$ is jointly concave in $(M, N) \in S_+^2$ (see Lieb (1973), Corollary 1.1) where S_+ is the space of positive semi-definite matrices so we have

$$\phi\left(\frac{1}{2}(\Sigma', \Sigma', I) + \frac{1}{2}(\Sigma_1, \Sigma_2, I)\right) \geq \frac{1}{2}\phi(\Sigma', \Sigma', I) + \frac{1}{2}\phi(\Sigma_1, \Sigma_2, I),$$

i.e.

$$\begin{aligned} Tr\left((\Sigma_1 + \Sigma')^{\frac{1}{2}}(\Sigma_2 + \Sigma')^{\frac{1}{2}}\right) =& Tr\left((\Sigma_1 + \Sigma')^{\frac{1}{4}}(\Sigma_2 + \Sigma')^{\frac{1}{2}}(\Sigma_1 + \Sigma')^{\frac{1}{4}}\right) \\ &\geq Tr\left(\Sigma'\right) + Tr\left(\Sigma_1^{\frac{1}{4}}\Sigma_2^{\frac{1}{2}}\Sigma_1^{\frac{1}{4}}\right) = Tr\left(\Sigma' + \Sigma_1^{\frac{1}{2}}\Sigma_2^{\frac{1}{2}}\right). \end{aligned}$$

Using this inequality and (18), we obtain that F satisfies **(SI)**. □

D.10.2 Projection

Similarly to the $d = 1$ case, using that $\mathrm{MW}^2_{F,\mathcal{M}^d}(\eta, \eta')$ is defined as the minimum over $w \in \Pi(\pi, \pi')$ of

$$\Lambda(w, \eta, \eta') = \sum_{\ell,k} w_{\ell k}\left(\sum_i (\mu_{\ell i} - \mu'_{ki})^2 + \sum_{i,j}(\Sigma_{\ell ij}^{1/2} - (\Sigma')_{kij}^{1/2})^2\right),$$

we can replace x and y in (15) with successively the means coefficients and then the covariance matrices coefficients of the two mixtures components to get,

$$\Lambda(w, \eta, \eta') \geq \sum_{\ell,k} w_{\ell k}\left(\sum_i\left(\mu_{\ell i} - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\mu_{\ell' i}\right)^2 + \sum_{i,j}\left(\Sigma_{\ell ij}^{1/2} - \sum_{\ell'}\frac{w_{\ell' k}}{\pi'_k}\Sigma_{\ell' ij}^{1/2}\right)^2\right).$$

Following the same methodology as in Theorem 6, we define $w^*(\eta)$ as a minimiser of the previous expression, $\tilde{\mu}_k(\eta) = \sum_{\ell'}\frac{w^*_{\ell' k}(\eta)}{\pi'_k}\mu_{\ell'}$, $\tilde{\Sigma}_k^{1/2}(\eta) = \sum_{\ell'}\frac{w^*_{\ell' k}(\eta)}{\pi'_k}\Sigma_{\ell'}^{1/2}$, $\tilde{\pi}_k(\eta) = \sum_{\ell'} w^*_{\ell' k}(\eta)$, and we get the following projection

$$\Pi^{w^*}_{\mathrm{MW}_{F,\mathcal{M}^d}}\eta = \sum_k \tilde{\pi}_k(\eta)\mathcal{N}(\tilde{\mu}_k(\eta), \tilde{\Sigma}_k(\eta)) \in \mathcal{M}_K.$$

As previously, we can replace the x_i's' and y_i's' in (16) by the means coefficients and covariance matrix coefficients of the two mixtures components to get

$$\mathrm{MW}^2_{F,\mathcal{M}^d}(\Pi^{w^*}_{\mathrm{MW}_{F,\mathcal{M}^d}}\eta, \Pi^{w^*}_{\mathrm{MW}_{F,\mathcal{M}^d}}\eta') \leq \mathrm{MW}^2_{F,\mathcal{M}^d}(\eta, \eta'),$$

thus $\Pi^{w^*}_{\mathrm{MW}_{F,\mathcal{M}^d}}$ is a non-expansion with respect to $\mathrm{MW}_{F,\mathcal{M}^d}$.

D.11 Expressions of MMD distances for GM

In the MMD case, Theorem 10 provides closed-form expressions for a number of kernel choices, when the compared distributions are mixtures or Gaussians.

Theorem 10. *Let* $X \sim \sum_{k=1}^{K_1} \pi_{1k}\mathcal{N}(\mu_{1k}, \sigma_{1k}^2)$ *and* $Y \sim \sum_{k=1}^{K_2} \pi_{2k}\ \mathcal{N}(\mu_{2k}, \sigma_{2k}^2)$*, two independent random variables distributed as GM. We have the following closed-form expressions:*

$$\begin{aligned}
\mathbb{E}[k_{\gamma_0,\mathrm{lap}}(X,Y)] &= \sum_{k,\ell} \pi_{1k}\pi_{2\ell} G\left(\gamma_0\tilde{\mu}_{k,\ell}, \gamma_0\tilde{\sigma}_{k,\ell}\right), \\
\mathbb{E}[k_{\gamma_1,\mathrm{rbf}}(X,Y)] &= \sum_{k,\ell} \pi_{1k}\pi_{2\ell} \frac{\gamma_1}{\sqrt{\gamma_1^2 + \tilde{\sigma}_{k,\ell}^2}} e^{-\frac{\tilde{\mu}_{k,\ell}^2}{2\left(\gamma_1^2 + \tilde{\sigma}_{k,\ell}^2\right)}}, \\
\mathbb{E}[k_{\mathrm{en}}(X,Y)] &= -\sum_{k,\ell} \pi_{1k}\pi_{2\ell}\tilde{\sigma}_{k,\ell} V\left(\frac{\tilde{\mu}_{k,\ell}}{\tilde{\sigma}_{k,\ell}}\right),
\end{aligned}$$

$$\textit{where } G\left(x,y\right) = F(x,y) + F(-x,y) \quad \textit{with } F(x,y) = e^{\frac{y^2}{2} - x} F_{\mathcal{N}(0,1)}\left(\frac{x}{y} - y\right),$$

$$V(x) = x(2F_{\mathcal{N}(0,1)}(x) - 1) + 2\mathcal{N}(x;0,1) \quad \textit{and } \tilde{\sigma}_{k,\ell}^2 = \sigma_{1k}^2 + \sigma_{2\ell}^2, \quad \tilde{\mu}_{k,\ell} = \mu_{1k} - \mu_{2\ell}.$$

Proof. For any kernel k,

$$\mathbb{E}[k(X,Y)] = \sum_{k=1}^{K_1}\sum_{\ell=1}^{K_2} \pi_{1k}\pi_{2\ell}\ \mathbb{E}[k(X_k, Y_\ell)],$$

where $X_k \sim \mathcal{N}(\mu_{1k}, \sigma_{1k}^2)$, $Y_\ell \sim \mathcal{N}(\mu_{2\ell}, \sigma_{2\ell}^2)$. We only need to compute expectations for two Gaussian variables.

Laplacian kernel. Since $X_k - Y_\ell \sim \mathcal{N}(\mu_{1k} - \mu_{2\ell}, \sigma_{1k}^2 + \sigma_{2\ell}^2)$, for all $\gamma_0 \geq 0$, we need to compute $\mathbb{E}[e^{-\gamma_0|Z|}]$ where Z is univariate Gaussian distributed. For $Z \sim \mathcal{N}(\mu, \sigma^2)$,

$$\begin{aligned}
\int_0^{+\infty} e^{-\gamma_0 x}\eta_Z(x)dx &= e^{\frac{(\gamma_0\sigma)^2}{2} - \mu\gamma_0} \int_0^{+\infty} \eta_Z(x + \gamma_0\sigma^2)dx \\
&= e^{\frac{(\gamma_0\sigma)^2}{2} - \mu\gamma_0} \int_{\gamma_0\sigma - \frac{\mu}{\sigma}}^{+\infty} \eta_{\frac{Z-\mu}{\sigma}}(x)dx \\
&= e^{\frac{(\gamma_0\sigma)^2}{2} - \mu\gamma_0} F_{\mathcal{N}(0,1)}\left(\frac{\mu}{\sigma} - \gamma_0\sigma\right).
\end{aligned}$$

From which, we deduce that

$$\begin{aligned}
\mathbb{E}[e^{-\gamma_0|Z|}] &= \int_{\mathbb{R}} e^{-\gamma_0|x|}\eta_Z(x)dx \\
&= \int_0^{+\infty} e^{-\gamma_0 x}(\eta_Z(x) + \eta_Z(-x))dx \\
&= \int_0^{+\infty} e^{-\gamma_0 x}(\eta_Z(x) + \eta_{-Z}(x))dx.
\end{aligned}$$

Hence,

$$\begin{aligned}
\mathbb{E}[e^{-\gamma_0|X_k - Y_\ell|}] =& e^{\frac{(\gamma_0\tilde{\sigma}_{k,\ell})^2}{2} - \gamma_0\tilde{\mu}_{k\ell}} F_{\mathcal{N}(0,1)}\left(\frac{\tilde{\mu}_{k\ell}}{\tilde{\sigma}_{k\ell}} - c\tilde{\sigma}_{k\ell}\right) \\
&+ e^{\frac{(\gamma_0\tilde{\sigma}_{k\ell})^2}{2} + \gamma_0\tilde{\mu}_{k\ell}} F_{\mathcal{N}(0,1)}\left(\frac{-\tilde{\mu}_{k\ell}}{\tilde{\sigma}_{k\ell}} - \gamma_0\tilde{\sigma}_{k\ell}\right).
\end{aligned}$$

Gaussian kernel. We also have

$$\begin{aligned}\mathbb{E}[e^{-\frac{|X_k-Y_\ell|^2}{2\gamma_1^2}}] =&\sqrt{2\pi}\gamma_1 \int_{\mathbb{R}} \mathcal{N}(x \mid 0, \gamma_1^2)\mathcal{N}(x \mid \tilde{\mu}_{k\ell}, \tilde{\sigma}_{k\ell}^2)dx \\ =&\sqrt{2\pi}\gamma_1 \mathcal{N}(\tilde{\mu}_{k\ell} \mid 0, \gamma_1^2 + \tilde{\sigma}_{k\ell}^2) \\ =&\frac{\gamma_1}{\sqrt{\gamma_1^2 + \tilde{\sigma}_{k\ell}^2}} e^{-\frac{\tilde{\mu}_{k\ell}^2}{2(\gamma_1^2+\tilde{\sigma}_{k\ell}^2)}}.\end{aligned}$$

Energy kernel. Using that

$$\mathbb{E}[|Z|] = \int_0^\infty (1 - F_{|Z|}(x))\, dx = \int_{-\infty}^0 (F_Z(x) + F_{-Z}(x))\, dx \tag{19}$$

and considering the standardization with $U = (Z - \mu)/\sigma \sim \mathcal{N}(0, 1)$, it comes

$$\mathbb{E}[|Z|] = \sigma \left(\int_{-\infty}^{-\mu/\sigma} F_U(x)\, dx + \int_{-\infty}^{\mu/\sigma} F_{-U}(x)\, dx \right) .$$

Integration by parts leads to

$$\begin{aligned}\int_{-\infty}^{u} F_U(x)\, dx &= u \int_{-\infty}^{u} \mathcal{N}(t; 0, 1)\, dt + \mathcal{N}(u; 0, 1) \\ &= uF_U(u) + \mathcal{N}(u; 0, 1) \\ \int_{-\infty}^{-u} F_U(x)\, dx &= -uF_U(-u) + \mathcal{N}(u; 0, 1) \\ &= -u(1 - F_U(u)) + \mathcal{N}(u; 0, 1)\end{aligned}$$

and to $\mathbb{E}[|Z|] = \sigma\, (u(2F_U(u) - 1) + 2\mathcal{N}(u; 0, 1)) = \sigma V(u)$, where $V(u)$ is as stated in the theorem. □

D.12 Equivalence between the Cramer and energy distance

Zhang (2023) implicitly provides a proof that the Cramer distance is proportional to $\mathrm{MMD}_{k_{\mathrm{en}}}$ that is different from the already well-known geometric proof cited by Nam et al. (2021). To highlight this, let us consider the following lemma.

Lemma 2. *Let X, Y be two independent random variables. We have the following formula,*

$$\forall r \in \mathbb{R}, \int_{\mathbb{R}} F_X(x)(1 - F_{r+Y}(x))dx = \int_{-\infty}^{r} F_{X-Y}(x)dx, \tag{20}$$

where F_X is the CDF of X.

Proof. We adapt here the proof used in the particular case of Gaussian mixtures in Zhang (2023). Indeed, we have

$$\begin{aligned}&\frac{\partial^2}{\partial r^2} \int_{\mathbb{R}} F_X(x)(1 - F_{Y+r}(x))dx = \frac{\partial}{\partial r} \int_{\mathbb{R}} \frac{\partial}{\partial r} \left(1 - F_Y\left(x - r\right)\right) F_X(x)dx \\ &= \frac{\partial}{\partial r} \int_{\mathbb{R}} \eta_Y(x)F_X(x + r)dx = \int_{\mathbb{R}} \eta_Y(x)\eta_X(x + r)dx \\ &= \int_{\mathbb{R}} \eta_{Y+r}(x)\eta_X(x)dx = (\eta_{-Y} * \eta_X)(r).\end{aligned}$$

Hence there exist constants C_1, C_2 independent of r such that

$$\int_{\mathbb{R}} F_X(x)(1 - F_{Y+r}(x))dx = C_1 + C_2 r + \int_{-\infty}^{r} \int_{-\infty}^{x} (\eta_{-Y} * \eta_X)(t)dtdx.$$

By taking the limit $r \to -\infty$, it comes $C_1 = C_2 = 0$ and thus

$$\int_{\mathbb{R}} F_X(x)(1 - F_{Y+r}(x))dx = \int_{-\infty}^{r} \int_{-\infty}^{x} (\eta_{-Y} * \eta_X)(t)dtdx = \int_{-\infty}^{r} F_{X-Y}(x)dx.$$

□

Using this lemma, we can prove that the Cramer distance C is proportional to $\mathrm{MMD}_{k_{\mathrm{en}}}$ if $d = 1$. The result is well-known but we provide a proof different from the classical proof of Szekely (2002).

Theorem 11. *Let P, Q two distributions, we have $C^2(P, Q) = \frac{1}{2}\mathrm{MMD}^2_{k_{\mathrm{en}}}(P, Q)$.*

Proof. The Cramer distance is defined by $C^2(P, Q) = \int_{\mathbb{R}} |F_P(x) - F_Q(x)|^2 dx$. Then,

$$\begin{aligned}\int_{\mathbb{R}} |F_P(x) - F_Q(x)|^2 dx = &\int_{\mathbb{R}} F_P(x)(1 - F_Q(x))dx + \int_{\mathbb{R}} F_Q(x)(1 - F_P(x))dx \\ &- \int_{\mathbb{R}} F_P(x)(1 - F_P(x))dx - \int_{\mathbb{R}} F_Q(x)(1 - F_Q(x))dx,\end{aligned}$$

For some independent variables $(X, \tilde{X}, Y, \tilde{Y})$ with $X, \tilde{X} \sim P$ and $Y, \tilde{Y} \sim Q$, Lemma 2 leads to

$$\int_{\mathbb{R}} |F_P(t) - F_Q(t)|^2 dt = \int_{-\infty}^{0} F_{X-Y}(x)dx + \int_{-\infty}^{0} F_{Y-X}(x)dx - \int_{-\infty}^{0} F_{X-\tilde{X}}(x)dx - \int_{-\infty}^{0} F_{Y-\tilde{Y}}(x)dx.$$

Using formula (19) and that $F_{X-\tilde{X}} = F_{\tilde{X}-X}$ and $F_{Y-\tilde{Y}} = F_{\tilde{Y}-Y}$ by symmetry, it follows,

$$2\int_{\mathbb{R}} |F_P(t) - F_Q(t)|^2 dt = 2\mathbb{E}[|X - Y|] - \mathbb{E}[|X - \tilde{X}|] - \mathbb{E}[|Y - \tilde{Y}|] = \mathrm{MMD}^2_{k_{\mathrm{en}}}(P, Q).$$

□

E Stochastic Approximation solution

Although most DRL solutions are based on a TD learning principle, the approximation of returns by Gaussian mixtures opens the way to other approaches to access the targeted fixed point. Stochastic approximation (SA) is a popular approach for solving fixed-point equations where the information is corrupted by noise. Such is the case in RL and this technique has theoretical advantages over TD-learning (Chen et al., 2020). More recently, it has shown tremendous results in DRL using Gaussian mixtures (Zhang et al., 2024). The principle is simpler than TD-learning. The online network is updated using a point-wise stochastic approximation,

$$\bar{\eta}_{t+1}(x, a) = (1 - \beta)\bar{\eta}_t(x, a) + \beta(r + T\gamma)_{\#}\bar{\eta}_t(x', a'),$$

with a transition sample (x, a, r, x', a'). As the right-hand term is a mixture of two distributions, it is natural to use mixtures as representations, as a mixture of mixtures is a mixture. However, it is impractical to use directly the resulting mixture since the number of components increases exponentially. The solution of Zhang et al. (2024) uses an EM estimation to approximate the resulting mixture by a fixed number of components. This is crucial in their approach which then optimizes a loss that can only be computed between mixtures with the same component numbers. One advantage of the metrics we propose is that this approximation is not necessary as they can handle mixtures of arbitrary sizes. It follows the possibility to design new simpler algorithms whose global principle is illustrated in Algorithm 2 (sg is the stopgrad operation). In practice, this would require refinements out of the scope of this paper.

F Experimental setup and hyperparameters

All experiments were run on our local cluster. The main experiments on Atari-5 were run on NVIDIA L40S GPUs. The ablation study on Atari-3 was carried out on NVIDIA A40 GPUs. Additional experiments

Algorithm 2: SA training for 1 episode

```
while environment is not terminated do
    Sample $a_t \sim \pi_{\varepsilon-\text{greedy}}(\cdot \mid x_t)$
    Get $x_{t+1} \sim P(\cdot \mid x_t, a_t)$ and $r_t \sim R(x_t, a_t)$
    $\eta_{\text{new}} \leftarrow (1-\beta)\text{sg}(\bar{\eta}_t(x_t, a_t))$
    $\eta_{\text{new}} \leftarrow \eta_{\text{new}} + \beta\text{sg}((R + T\gamma)_{\#}\bar{\eta}_t(x_{t+1}, a_{t+1}))$
    Minimize $D(\eta_t(x_t, a_t), \eta_{\text{new}})$ over network parameters
    $t \leftarrow t + 1$
```

shown in Figure 6 were run on NVIDIA V100 GPUs. We mostly used the Dopamine (Castro et al., 2018) hyperparameters recommandations (Table 3) but we had to use gradient clipping to stabilize the training. To guarantee fair comparison, we used the same hyperparameters for all methods for each environment and we selected the gradient clipping by selecting for each environment the highest possible giving a stable training for all methods. This leads in our case to gradclip = 100 for Asterix and Bowling, and gradclip = 1 for Qbert. As regards the GM representations, the number of components was set to $K = 5$ in the Atari-5 experiments, while K was varying in $\{1, 3, 5, 16\}$ in the ablation study. To compute metrics, the MW_2 was implemented using the POT package (Flamary et al., 2021) and for MMD computations, we used the same setting as Nguyen-Tang et al. (2020).

Table 3: Hyperparameters used in experiments

Hyperparameter	**Setting**
Sticky actions probability	0.25 (stochastic case)
Discount factor (γ)	0.99
Frames stacked	4
Mini-batch size ($\mathcal{B}$)	32
Replay memory size	10^5
Online network update rate (λ)	4 steps
Target network update rate (τ)	10^4 steps
Initial exploration (ϵ)	1
Exploration decay rate	10^{-2}
Exploration decay period	2.5×10^5 steps
Test exploration	10^{-3}
Environment steps per iteration	2.5×10^5 steps
Starting step	5×10^4 steps
Adam hyperparameters	
β_1 decay	0.9
β_2 decay	0.999
Eps	$\frac{10^{-2}}{\mathcal{B}}$
Learning rate	5×10^{-5}

G Additional experiments

Table 4 and Figure 5 provide results for an ablation study, varying the number of mixture components $K \in \{1, 3, 5, 16\}$ in the deterministic case with the Atari-3 subset of 3 games proposed in Aitchison et al. (2023). Figure 6 provides additional illustrations on 3 other Atari games, in a deterministic setting, plus one of the games in a stochastic setting similar to that of Section 8.

Table 4: Ablation over the number of components K, in the deterministic case with metric MW_2. Normalized scores on Atari-3 using the weights given in Aitchison et al. (2023), for varying K.

Environment	$K = 1$	$K = 3$	$K = 5$	$K = 16$
Deterministic	43.80 ± 3.63	44.40 ± 13.20	$\mathbf{48.37 \pm 13.36}$	31.82 ± 19.98

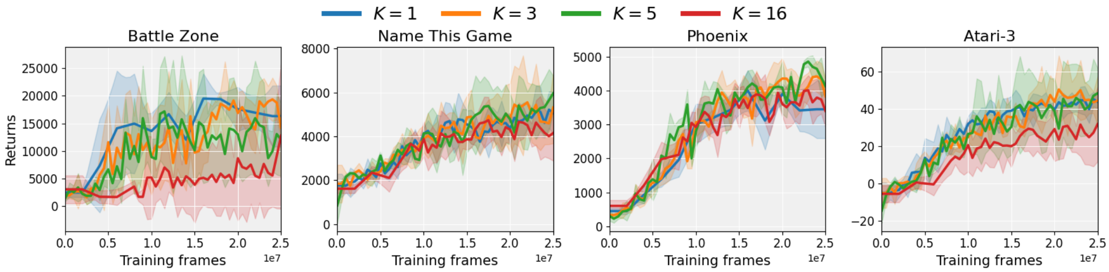
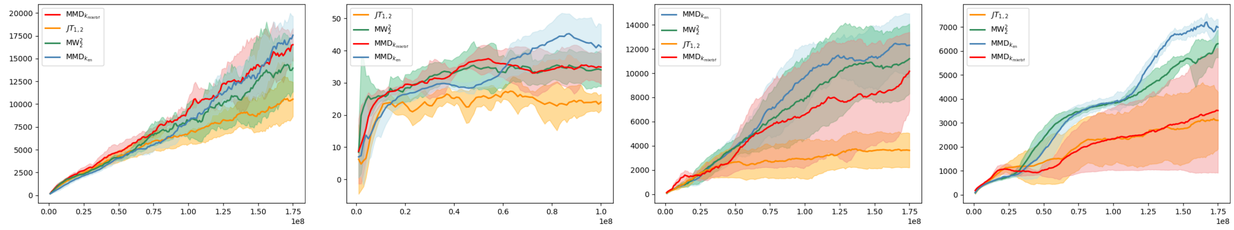


Figure 5: Ablation over the number of components K, using the Atari-3 games in the deterministic case with metric MW_2. Moving average return, over 500k frames, with respect to the number of training frames. Curves are averaged over 5 runs with shaded areas representing standard deviations.



(a) Asterix

(b) Bowling

(c) Qbert

(d) Qbert stochastic

Figure 6: Comparison of MW_2(green), $\text{JT}_{1,2}$(yellow), $\text{MMD}_{k_{en}}$(blue) and $\text{MMD}_{k_{mixrbf}}$(red) metrics for Atari games in deterministic (a,b,c) and stochastic (d) environments. Moving average return, over 500k frames, with respect to the number of training frames. Curves are averaged over 5 runs with shaded areas representing standard deviations.

H Computation times

Tables 5 and 6 present respectively computation times for the Atari-5 experiments and the ablation study on K for the Atari-3 games. Computation times tend to increase with K.

Table 5: Average computation overheads and standard deviations (in hours) on Atari-5 for the different compared metrics in deterministic and stochastic environments. The shortest times are indicated in bold characters.

Environment	$\mathbf{JT}_{1,2}$	$\mathbf{MW}_2$	$\mathbf{MMD}_{\mathrm{k_{en}}}$	$\mathbf{MMD}_{\mathrm{k_{mixrbf}}}$
Stochastic	46.12 ± 0.27	166.45 ± 0.42	46.38 ± 0.16	$\mathbf{45.71 \pm 0.06}$
Deterministic	44.45 ± 0.57	166.25 ± 0.65	46.29 ± 0.47	$\mathbf{43.68 \pm 0.33}$

Table 6: Average computation overheads and standard deviations (in hours) on Atari-3 when using the MW_2 metric in a deterministic environment and varying the number of mixture components K. The shortest times are indicated in bold characters.

Environment	$K = 1$	$K = 3$	$K = 5$
Deterministic	$\mathbf{16.02 \pm 0.18}$	62.05 ± 0.31	135.46 ± 0.53