Equivalence Between Wasserstein and Value-Aware Loss for Model-based Reinforcement Learning

Kavosh Asadi [1] **Evan Cater** [1] **Dipendra Misra** [2] **Michael L. Littman** [1]

Abstract

Learning a generative model is a key component of model-based reinforcement learning. Though learning a good model in the tabular setting is a simple task, learning a useful model in the approximate setting is challenging. In this context, an important question is the loss function used for model learning as varying the loss function can have a remarkable impact on effectiveness of planning. Recently Farahmand et al. (2017) proposed a value-aware model learning (VAML) objective that captures the structure of value function during model learning. Using tools from Asadi et al. (2018), we show that minimizing the VAML objective is in fact equivalent to minimizing the Wasserstein metric. This equivalence improves our understanding of value-aware models, and also creates a theoretical foundation for applications of Wasserstein in model-based reinforcement learning.

1. Introduction

The model-based approach to reinforcement learning consists of learning an internal model of the environment and planning with the learned model (Sutton & Barto, 1998). The main promise of the model-based approach is data-efficiency: the ability to perform policy improvements with a relatively small number of environmental interactions.

Although the model-based approach is well-understood in the tabular case (Kaelbling et al., 1996; Sutton & Barto, 1998), the extension to approximate setting is difficult. Models usually have non-zero generalization error due to limited training samples. Moreover, the model learning problem can be unrelizable, leading to an imperfect model with irreducible error (Ross & Bagnell, 2012; Talvitie, 2014). Sometimes referred to as the compounding error phenomenon, it has been shown that such small modeling errors can also compound after multiple steps and degrade the policy learned using the model (Talvitie, 2014; Venkatraman et al., 2015; Asadi et al., 2018).

On way of addressing this problem is by learning a model that is tailored to the specific planning algorithm we intend to use. That is, even though the model is imperfect, it is useful for the planning algorithm that is going to leverage it. To this end, Farahmand et al. (2017) proposed an objective function for model-based RL that captures the structure of value function during model learning to ensure that the model is useful for Value Iteration. Learning a model using this loss, known as value-aware model learning (VAML) loss, empirically improved upon a model learned using maximum-likelihood objective, thus providing a promising direction for learning useful models in the approximate setting.

More specifically, VAML minimizes the maximum Bellman error given the learned model, MDP dynamics, and an arbitrary space of value functions. As we will show, computing the Wasserstein metric involves a similar maximization problem, but over a space of Lipschitz functions. Under certain assumptions, we prove that the value function of an MDP is Lipschitz. Therefore, minimizing the VAML objective is in fact equivalent to minimizing Wasserstein.

2. Background

2.1. MDPs

We consider the Markov decision process (MDP) setting in which the RL problem is formulated by the tuple $\langle \mathcal{S}, \mathcal{A}, R, T, \gamma \rangle$. Here, $\mathcal{S}$ denotes a state space and $\mathcal{A}$ denotes an action set. The functions $R : S \times A \to \mathbb{R}$ and T: $\mathcal{S} \times \mathcal{A} \to \Pr(\mathcal{S})$ denote the reward and transition dynamics. Finally $\gamma \in [0, 1)$ is the discount rate.

[1]Department of Computer Science, Brown University, Providence, USA [2]Department of Computer Science, Cornell Tech, New York, USA. Correspondence to: Kavosh asadi <kavosh@brown.edu>.

Accepted at the FAIM workshop "Prediction and Generative Modeling in Reinforcement Learning", Stockholm, Sweden, 2018.

2.2. Lipschitz Continuity

We make use of the notion of "smoothness" of a function as quantified below.

Definition 1. *Given two metric spaces (M_1, d_1) and (M_2, d_2) consisting of a space and a distance metric, a function $f : M_1 \mapsto M_2$ is* Lipschitz continuous *(sometimes simply Lipschitz) if the Lipschitz constant, defined as*

$$K_{d_1,d_2}(f) := \sup_{s_1 \in \mathcal{S}, s_2 \in \mathcal{S}} \frac{d_2\big(f(s_1), f(s_2)\big)}{d_1(s_1, s_2)} , \qquad (1)$$

is finite.

Equivalently, for a Lipschitz f,

$$\forall s_1, \forall s_2 \quad d_2\big(f(s_1), f(s_2)\big) \leq K_{d_1,d_2}(f)\, d_1(s_1, s_2) \ .$$

Note that the input and output of f can generally be scalars, vectors, or probability distributions. A Lipschitz function f is called a *non-expansion* when $K_{d_1,d_2}(f) = 1$ and a *contraction* when $K_{d_1,d_2}(f) < 1$. We also define Lipschitz continuity over a subset of inputs:

Definition 2. *A function $f : M_1 \times \mathcal{A} \mapsto M_2$ is* uniformly Lipschitz continuous *in $\mathcal{A}$ if*

$$K^{\mathcal{A}}_{d_1,d_2}(f) := \sup_{a \in \mathcal{A}} \sup_{s_1,s_2} \frac{d_2\big(f(s_1, a), f(s_2, a)\big)}{d_1(s_1, s_2)} , \qquad (2)$$

is finite.

Note that the metric d_1 is still defined only on M_1. Below we also present two useful lemmas.

Lemma 1. *(Composition Lemma) Define three metric spaces (M_1, d_1), (M_2, d_2), and (M_3, d_3). Define Lipschitz functions $f : M_2 \mapsto M_3$ and $g : M_1 \mapsto M_2$ with constants $K_{d_2,d_3}(f)$ and $K_{d_1,d_2}(g)$. Then, $h : f \circ g : M_1 \mapsto M_3$ is Lipschitz with constant $K_{d_1,d_3}(h) \leq K_{d_2,d_3}(f) K_{d_1,d_2}(g)$.*

Proof.

$$\begin{aligned}
& K_{d_1,d_3}(h) \\
= & \sup_{s1,s_2} \frac{d_3\Big(f\big(g(s_1)\big), f\big(g(s_2)\big)\Big)}{d_1(s_1, s_2)} \\
= & \sup_{s_1,s_2} \frac{d_2\big(g(s_1), g(s_2)\big)}{d_1(s_1, s_2)} \frac{d_3\Big(f\big(g(s_1)\big), f\big(g(s_2)\big)\Big)}{d_2\big(g(s_1), g(s_2)\big)} \\
\leq & \sup_{s_1,s_2} \frac{d_2\big(g(s_1), g(s_2)\big)}{d_1(s_1, s_2)} \sup_{s_1,s_2} \frac{d_3\big(f(s_1), f(s_2)\big)}{d_2(s_1, s_2)} \\
= & K_{d_1,d_2}(g) K_{d_2,d_3}(f).
\end{aligned}$$

□

Lemma 2. *(Summation Lemma) Define two vector spaces $(M_1, \|\|)$ and $(M_2, \|\|)$. Define Lipschitz functions $f : M_1 \mapsto M_2$ and $g : M_1 \mapsto M_2$ with constants $K_{\|\|,\|\|}(f)$ and $K_{\|\|,\|\|}(g)$. Then, $h : f + g : M_1 \mapsto M_2$ is Lipschitz with constant $K_{\|\|,\|\|}(h) \leq K_{\|\|,\|\|}(f) + K_{\|\|,\|\|}(g)$.*

Proof.

$$\begin{aligned}
K_{d_1,d_2}(h) \quad := \quad & \sup_{s_1,s_2} \frac{\|f(s_2) + g(s_2) - f(s_1) - g(s_1)\|}{\|s_2 - s_1\|} \\
\leq \quad & \sup_{s_1,s_2} \frac{\|f(s_2) - f(s_1)\|}{\|s_2 - s_1\|} + \frac{\|g(s_2) - g(s_1)\|}{\|s_2 - s_1\|} \\
\leq \quad & \sup_{s_1,s_2} \frac{\|f(s_2) - f(s_1)\|}{\|s_2 - s_1\|} \\
+ \quad & \sup_{s_1,s_2} \frac{\|g(s_2) - g(s_1)\|}{\|s_2 - s_1\|} \\
= \quad & K_{\|\|,\|\|}(f) + K_{\|\|,\|\|}(g)
\end{aligned}$$

□

2.3. Distance Between Distributions

We require a notion of difference between two distributions quantified below.

Definition 3. *Given a metric space (M, d) and the set $\mathbb{P}(M)$ of all probability measures on M, the* Wasserstein metric *(or the 1st Kantorovic metric) between two probability distributions μ_1 and μ_2 in $\mathbb{P}(M)$ is defined as*

$$W(\mu_1, \mu_2) := \inf_{j \in \Lambda} \int\int j(s_1, s_2) d(s_1, s_2) ds_2 \, ds_1 \ , \quad (3)$$

where Λ denotes the collection of all joint distributions j on $M \times M$ with marginals μ_1 and μ_2 (Vaserstein, 1969).

Wasserstein is linked to Lipschitz continuity using duality:

$$W(\mu_1, \mu_2) = \sup_{f : K_{d,d_{\mathbb{R}}}(f) \leq 1} \int \big(\mu_1(s) - \mu_2(s)\big) f(s) ds \ . \quad (4)$$

This equivalence is known as Kantorovich-Rubinstein duality (Kantorovich & Rubinstein, 1958; Villani, 2008). Sometimes referred to as "Earth Mover's distance", Wasserstein has recently become popular in machine learning, namely in the context of generative adversarial networks (Arjovsky et al., 2017) and value distributions in reinforcement learning (Bellemare et al., 2017). We also define Kullback Leibler divergence (simply KL) as an alternative measure of difference between two distributions:

$$KL(\mu_1 \,||\, \mu_2) := \int \mu_1(s) \log \frac{\mu_1(s)}{\mu_2(s)} ds \ .$$

3. Value-Aware Model Learning (VAML) Loss

The basic idea behind VAML (Farahmand et al., 2017) is to learn a model tailored to the planning algorithm that intends to use it. Since Bellman equations (Bellman, 1957) are in the core of many RL algorithms (Sutton & Barto, 1998), we assume that the planner uses the following Bellman equation:

$$Q(s,a) = R(s,a) + \gamma \int T(s'|s,a) f\big(Q(s',.)\big) ds' \,,$$

where f can generally be any arbitrary operator (Littman & Szepesvári, 1996) such as max. We also define:

$$v(s') := f\big(Q(s',.)\big) \,.$$

A good model $\widehat{T}$ could then be thought of as the one that minimizes the error:

$$\begin{aligned} l(T,\widehat{T})(s,a) &= R(s,a) + \gamma \int T(s'|s,a) v(s') ds' \\ &- R(s,a) - \gamma \int \widehat{T}(s'|s,a) v(s') ds' \\ &= \gamma \int \big(T(s'|s,a) - \widehat{T}(s'|s,a)\big) v(s') ds' \end{aligned}$$

Note that minimizing this objective requires access to the value function in the first place, but we can obviate this need by leveraging Holder's inequality:

$$\begin{aligned} l(\widehat{T},T)(s,a) &= \gamma \int \big(T(s'|s,a) - \widehat{T}(s'|s,a)\big) v(s') ds' \\ &\leq \gamma \left\| T(s'|s,a) - \widehat{T}(s'|s,a) \right\|_1 \|v\|_\infty \end{aligned}$$

Further, we can use Pinsker's inequality to write:

$$\left\| T(\cdot|s,a) - \widehat{T}(\cdot|s,a) \right\|_1 \leq \sqrt{2KL\big(T(\cdot|s,a) || \widehat{T}(\cdot|s,a)\big)} \,.$$

This justifies the use of maximum likelihood estimation for model learning, a common practice in model-based RL (Bagnell & Schneider, 2001; Abbeel et al., 2006; Agostini & Celaya, 2010), since maximum likelihood estimation is equivalent to empirical KL minimization.

However, there exists a major drawback with the KL objective, namely that it ignores the structure of the value function during model learning. As a simple example, if the value function is constant through the state-space, any randomly chosen model $\widehat{T}$ will, in fact, yield zero Bellman error. However, a model learning algorithm that ignores the structure of value function can potentially require many samples to provide any guarantee about the performance of learned policy.

Consider the objective function $l(T,\widehat{T})$, and notice again that v itself is not known so we cannot directly optimize for this objective. Farahmand et al. (2017) proposed to search for a model that results in lowest error given all possible value functions belonging to a specific class:

$$L(T,\hat{T})(s,a) = \sup_{v \in \mathcal{F}} \left| \int \big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big) v(s') ds' \right|^2 \tag{5}$$

Note that minimizing this objective is shown to be tractable if, for example, $\mathcal{F}$ is restricted to the class of exponential functions. Observe that the VAML objective (5) is similar to the dual of Wasserstein (4), but the main difference is the space of value functions. In the next section we show that even the space of value functions are the same under certain conditions.

4. Lipschitz Generalized Value Iteration

We show that solving for a class of Bellman equations yields a Lipschitz value function. Our proof is in the context of GVI (Littman & Szepesvári, 1996), which defines Value Iteration (Bellman, 1957) with arbitrary backup operators. We make use of the following lemmas.

Lemma 3. *Given a non-expansion $f : \mathcal{S} \mapsto \mathbb{R}$:*

$$K^{\mathcal{A}}_{d_\mathcal{S},d_\mathbb{R}}\Big(\int T(s'|s,a) f(s') ds' \Big) \leq K^{\mathcal{A}}_{d_\mathcal{S},W}\big(T\big) \,.$$

Proof. Starting from the definition, we write:

$$\begin{aligned} &K^{\mathcal{A}}_{d_\mathcal{S},d_\mathbb{R}}\Big(\int T(s'|s,a) f(s') ds' \Big) \\ &= \sup_a \sup_{s_1,s_2} \frac{\left| \int \big(T(s'|s_1,a) - T(s'|s_2,a)\big) f(s') ds' \right|}{d(s_1,s_2)} \\ &\leq \sup_a \sup_{s_1,s_2} \frac{\left| \sup_g \int \big(T(s'|s_1,a) - T(s'|s_2,a)\big) g(s') ds' \right|}{d(s_1,s_2)} \\ &\text{(where } K_{d_\mathcal{S},d_\mathbb{R}}(g) \leq 1\text{)} \\ &= \sup_a \sup_{s_1,s_2} \frac{\sup_g \int \big(T(s'|s_1,a) - T(s'|s_2,a)\big) g(s') ds'}{d(s_1,s_2)} \\ &= \sup_a \sup_{s_1,s_2} \frac{W\big(T(\cdot|s_1,a), T(\cdot|s_2,a)\big)}{d(s_1,s_2)} = K^{\mathcal{A}}_{d_\mathcal{S},W}(T) \,. \end{aligned}$$

□

Lemma 4. *The following operators are non-expansion ($K_{\|\cdot\|_\infty,d_R}(\cdot) = 1$):*

1. $max(x)$, $mean(x)$
2. $\epsilon\text{-}greedy(x) := \epsilon\, mean(x) + (1-\epsilon) max(x)$
3. $mm_\beta(x) := \frac{\log \frac{\sum_i e^{\beta x_i}}{n}}{\beta}$

Proof. 1 is proven by Littman & Szepesvári (1996). 2 follows from 1: (metrics not shown for brevity)

$$\begin{aligned} K(\epsilon\text{-greedy(x)}) &= K\big(\epsilon \text{ mean}(x) + (1-\epsilon)\text{max}(x)\big) \\ &\leq \epsilon K\big(\text{mean}(x)\big) + (1-\epsilon)K\big(\text{max}(x)\big) \\ &= 1 \end{aligned}$$

Finally, 3 is proven multiple times in the literature. (Asadi & Littman, 2017; Nachum et al., 2017; Neu et al., 2017) □

Algorithm 1 GVI algorithm

Input: initial $\widehat{Q}(s,a)$, δ, and choose an operator f
repeat
  diff $\leftarrow 0$
  for each $s \in \mathcal{S}$ **do**
    for each $a \in \mathcal{A}$ **do**
      $Q_{copy} \leftarrow \widehat{Q}(s,a)$
      $\widehat{Q}(s,a) \leftarrow R(s,a) + \gamma \int T(s' \mid s,a) f\big(\widehat{Q}(s',\cdot)\big) ds'$
      diff $\leftarrow \max\big\{\text{diff}, |Q_{copy} - Q(s,a)|\big\}$
    end for
  end for
until diff $< \delta$

We now present the main result of this paper.

Theorem. *For any choice of backup operator f outlined in Lemma 4, GVI computes a value function with a Lipschitz constant bounded by $\frac{K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R)}{1-\gamma K_{d_{\mathcal{S}},W}(T)}$ if $\gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T) < 1$.*

Proof. From Algorithm 1, in the nth round of GVI updates we have:

$$\widehat{Q}_{n+1}(s,a) \leftarrow R(s,a) + \gamma \int T(s' \mid s,a) f\big(\widehat{Q}_n(s',\cdot)\big) ds'.$$

First observe that:

$$\begin{aligned} & K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(\widehat{Q}_{n+1}) \\ & \big(\text{due to Summation Lemma (2)}\big) \\ & \leq K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R) + \gamma K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}\Big(\int T(s' \mid s,a) f\big(\widehat{Q}_n(s',\cdot)\big) ds'\Big) \\ & \big(\text{due to Lemma (3)}\big) \\ & \leq K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R) + \gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T)\, K_{d_{\mathcal{S}},\mathbb{R}}\Big(f\big(\widehat{Q}_n(s,\cdot)\big)\Big) \\ & \big(\text{due to Composition Lemma (1)}\big) \\ & \leq K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R) + \gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T) K_{\|\cdot\|_\infty,d_{\mathbb{R}}}(f) K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{Q}_n) \\ & \big(\text{due to Lemma (4), the non-expansion property of } f\big) \\ & = K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R) + \gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T) K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{Q}_n) \end{aligned}$$

Equivalently:

$$\begin{aligned} K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(\widehat{Q}_{n+1}) &\leq K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(R) \sum_{i=0}^{n} \big(\gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T)\big)^i \\ &+ \big(\gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T)\big)^n K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{Q}_0)\,. \end{aligned}$$

By computing the limit of both sides, we get:

$$\begin{aligned} \lim_{n\to\infty} K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{Q}_n) &\leq \lim_{n\to\infty} K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(R) \sum_{i=0}^{n} \big(\gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T)\big)^i \\ &+ \lim_{n\to\infty} \big(\gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T)\big)^n K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{Q}_0) \\ &= \frac{K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R)}{1-\gamma K_{d_{\mathcal{S}},W}(T)} + 0\,, \end{aligned}$$

where we used the fact that

$$\lim_{n\to\infty} \big(\gamma K^{\mathcal{A}}_{d_{\mathcal{S}},W}(T)\big)^n = 0\,.$$

This concludes the proof. □

Now notice that as defined earlier:

$$\widehat{V}_n(s) := f\big(\widehat{Q}_n(s,\cdot)\big)\,,$$

so as a relevant corollary of our theorem we get:

$$\begin{aligned} K_{d_{\mathcal{S}},d_{\mathbb{R}}}\big(v(s)\big) &= \lim_{n\to\infty} K_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{V}_n) \\ &= \lim_{n\to\infty} K_{d_{\mathcal{S}},d_{\mathbb{R}}}\Big(f\big(\widehat{Q}_n(s,\cdot)\big)\Big) \\ &\leq \lim_{n\to\infty} K^{\mathcal{A}}_{d_{\mathcal{S}},d_{\mathbb{R}}}(\widehat{Q}_n) \\ &\leq \frac{K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R)}{1-\gamma K_{d_{\mathcal{S}},W}(T)}\,. \end{aligned}$$

That is, solving for the fixed point of this general class of Bellman equations results in a Lipschitz state-value function.

5. Equivalence Between VAML and Wasserstein

We now show the main claim of the paper, namely that minimzing for the VAML objective is the same as minimizing the Wasserstein metric.

Consider again the VAML objective:

$$L(T,\hat{T})(s,a) = \sup_{v\in\mathcal{F}} \Big| \int \big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big) v(s') ds' \Big|^2$$

where $\mathcal{F}$ can generally be any class of functions. From our theorem, however, the space of value functions $\mathcal{F}$ should be restricted to Lipschitz functions. Moreover, it is easy to design an MDP and a policy such that a desired Lipschitz value function is attained.

This space $\mathcal{L}_C$ can then be defined as follows:

$$\mathcal{L}_C = \{f : K_{d_{\mathcal{S}},d_{\mathbb{R}}}(f) \leq C\} ,$$

where

$$C = \frac{K^{\mathcal{A}}_{d_{\mathcal{S}},d_R}(R)}{1 - \gamma K_{d_{\mathcal{S}},W}(T)} .$$

So we can rewrite the VAML objective L as follows:

$$L\big(T,\hat{T}\big)(s,a) = \sup_{f\in\mathcal{L}_C} \Big| \int f(s)\big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big) ds' \Big|^2$$

$$= \sup_{f\in\mathcal{L}_C} \Big| \int C\frac{f(s)}{C}\big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big)\, ds' \Big|^2$$

$$= C^2 \sup_{g\in\mathcal{L}_1} \Big| \int g(s)\big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big)\, ds' \Big|^2 .$$

It is clear that a function g that maximizes the Kantorovich-Rubinstein dual form:

$$\sup_{g\in\mathcal{L}_1} \int g(s)\big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big)\, ds'$$

$$:= W(T(\cdot|s,a), \widehat{T}(\cdot|s,a)) ,$$

will also maximize:

$$L\big(T,\hat{T}\big)(s,a) = \Big| \int g(s)\big(T(s' \mid s,a) - \widehat{T}(s' \mid s,a)\big)\, ds' \Big|^2 .$$

This is due to the fact that $\forall g \in \mathcal{L}_1 \Rightarrow -g \in \mathcal{L}_1$ and so computing absolute value or squaring the term will not change $\arg\max$ in this case.

As a result:

$$L\big(T,\widehat{T}\big)(s,a) = \Big(C\, W\big(T(\cdot|s,a), \widehat{T}(\cdot|s,a)\big)\Big)^2 .$$

This highlights a nice property of Wasserstein, namely that minimizing this metric yields a value-aware model.

6. Conclusion and Future Work

We showed that the value function of an MDP is Lipschitz. This result enabled us to draw a connection between value-aware model-based reinforcement learning and the Wassertein metric.

We hypothesize that the value function is Lipschitz in a more general sense, and so, further investigation of Lipschitz continuity of value functions should be interesting on its own. The second interesting direction relates to design of practical model-learning algorithms that can minimize Wasserstein. Two promising directions are the use of generative adversarial networks (Goodfellow et al., 2014; Arjovsky et al., 2017) or approximations such as entropic regularization (Frogner et al., 2015). We leave these two directions for future work.

References

Abbeel, Pieter, Quigley, Morgan, and Ng, Andrew Y. Using inaccurate models in reinforcement learning. In *Proceedings of the 23rd international conference on Machine learning*, pp. 1–8. ACM, 2006.

Agostini, Alejandro and Celaya, Enric. Reinforcement learning with a gaussian mixture model. In *Neural Networks (IJCNN), The 2010 International Joint Conference on*, pp. 1–8. IEEE, 2010.

Arjovsky, Martin, Chintala, Soumith, and Bottou, Léon. Wasserstein generative adversarial networks. In *International Conference on Machine Learning*, pp. 214–223, 2017.

Asadi, Kavosh and Littman, Michael L. An alternative softmax operator for reinforcement learning. In *Proceedings of the 34th International Conference on Machine Learning*, pp. 243–252, 2017.

Asadi, Kavosh, Misra, Dipendra, and Littman, Michael L. Lipschitz continuity in model-based reinforcement learning. *arXiv preprint arXiv:1804.07193*, 2018.

Bagnell, J Andrew and Schneider, Jeff G. Autonomous helicopter control using reinforcement learning policy search methods. In *Robotics and Automation, 2001. Proceedings 2001 ICRA. IEEE International Conference on*, volume 2, pp. 1615–1620. IEEE, 2001.

Bellemare, Marc G, Dabney, Will, and Munos, Rémi. A distributional perspective on reinforcement learning. In *International Conference on Machine Learning*, pp. 449–458, 2017.

Bellman, Richard. A markovian decision process. *Journal of Mathematics and Mechanics*, pp. 679–684, 1957.

Farahmand, Amir-Massoud, Barreto, Andre, and Nikovski, Daniel. Value-Aware Loss Function for Model-based Reinforcement Learning. In *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics*, pp. 1486–1494, 2017.

Frogner, Charlie, Zhang, Chiyuan, Mobahi, Hossein, Araya, Mauricio, and Poggio, Tomaso A. Learning with a wasserstein loss. In *Advances in Neural Information Processing Systems*, pp. 2053–2061, 2015.

Goodfellow, Ian, Pouget-Abadie, Jean, Mirza, Mehdi, Xu, Bing, Warde-Farley, David, Ozair, Sherjil, Courville, Aaron, and Bengio, Yoshua. Generative adversarial nets. In *Advances in neural information processing systems*, pp. 2672–2680, 2014.

Kaelbling, Leslie Pack, Littman, Michael L., and Moore, Andrew W. Reinforcement learning: A survey. *J. Artif. Intell. Res.*, 4:237–285, 1996.

Kantorovich, Leonid Vasilevich and Rubinstein, G Sh. On a space of completely additive functions. *Vestnik Leningrad. Univ*, 13(7):52–59, 1958.

Littman, Michael L. and Szepesvári, Csaba. A generalized reinforcement-learning model: Convergence and applications. In *Proceedings of the 13th International Conference on Machine Learning*, pp. 310–318, 1996.

Nachum, Ofir, Norouzi, Mohammad, Xu, Kelvin, and Schuurmans, Dale. Bridging the gap between value and policy based reinforcement learning. *arXiv preprint arXiv:1702.08892*, 2017.

Neu, Gergely, Jonsson, Anders, and Gómez, Vicenç. A unified view of entropy-regularized Markov decision processes. *arXiv preprint arXiv:1705.07798*, 2017.

Ross, Stéphane and Bagnell, Drew. Agnostic system identification for model-based reinforcement learning. In *Proceedings of the 29th International Conference on Machine Learning, ICML 2012, Edinburgh, Scotland, UK, June 26 - July 1, 2012*, 2012.

Sutton, Richard S. and Barto, Andrew G. *Reinforcement Learning: An Introduction*. The MIT Press, 1998.

Talvitie, Erik. Model regularization for stable sample rollouts. In *Proceedings of the Thirtieth Conference on Uncertainty in Artificial Intelligence, UAI 2014, Quebec City, Quebec, Canada, July 23-27, 2014*, pp. 780–789, 2014.

Vaserstein, Leonid Nisonovich. Markov processes over denumerable products of spaces, describing large systems of automata. *Problemy Peredachi Informatsii*, 5(3):64–72, 1969.

Venkatraman, Arun, Hebert, Martial, and Bagnell, J Andrew. Improving multi-step prediction of learned time series models. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, January 25-30, 2015, Austin, Texas, USA.*, 2015.

Villani, Cédric. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.