Value Iteration Approximations using Reproducing Kernel Hilbert Spaces CS 598 - Statistical RL

Christian Howard
howard28@illinois.edu

Abstract

The Bellman equation for value functions and Markov Decision Processes (MDPs) are introduced, illustrating a fundamental dilemma in Reinforcement Learning when there is a lack of explicit knowledge of transition probabilities. This dilemma is used to motivate the need for methods that can efficiently and accurately approximate expectation integrals using sampled data, motivating the use of more recent methods based on Reproducing Kernel Hilbert Spaces (RKHS). RKHS are then introduced, along with the necessary functional analysis background, emphasizing the definitions and properties of the kernel functions. A discussion on how to model various probability distributions using RKHS follows, using measure theory and integral operators to represent and prove various useful properties. Mean maps are introduced and proofs for marginal approximations using mean maps and empirical samples are performed. Conditional mean maps are then discussed and a compact form for the empirical mean map is derived, finishing with results found in [1] for the convergence properties as dataset size grows. This leads into value iteration approximations using RKHS, where we start by defining MDPs, the Bellman equation, and the Bellman operator. Using these definitions and work on conditional mean maps, we show how to construct a conditional mean map that can approximate the expectation operator found in the Bellman operator and discuss some convergence results found in [2] using the resulting approximate Bellman operator in a value iteration algorithm scheme.

Contents

1 Motivation

Within the context of Reinforcement Learning, we face the burden of finding an optimal policy that can be used to obtain an optimal amount of value with respect to some Markov Decision Process (MDP). We encapsulate this value using the Bellman equation defined below

$$V^{\pi}(x) = R(x, \pi(x)) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,\pi(x))} \left[V^{\pi}(x') \right]$$

where V^{π} is the value function induced by some chosen policy π, subject to the dynamics $P(X'|X, A)$, reward function R, state space $\mathcal{X}$, action space $\mathcal{A}$, and discount factor γ. In an ideal environment, we have all of this information readily available and our state and action spaces are not overly large, so we can stick to tabular models that can be efficiently found using dynamic programming [3]. Unfortunately, in many problems we do not have sufficient knowledge about our dynamics defined by the transition probabilities $P(X'|X, A)$. This lack of information is crucial to being able to perform any sort of dynamic programming approach to Reinforcement Learning, forcing us to look elsewhere for tackling our problem. This is a sad result because dynamic programming approaches have the capability to obtain better optimal policies, ensuring global optimums in some cases, relative to more model-free approaches that tend to converge to local optimum.

Fortunately, there are some methods that allow use to use sampled data to estimate transition probabilities and in turn approximate the Bellman equation. Of course, even if you can collect enough data to sufficiently estimate these probabilities, computation of the expectation integral can be computationally intractable for high dimensional state spaces. This write up discusses developments of a more recent approach that uses Reproducing Kernel Hilbert Spaces to construct efficient expectation approximations that can be used to construct an efficient approximate Bellman equation for use in an approximate dynamic programming approach to value iteration.

2 Fundamentals of Reproducing Kernel Hilbert Spaces

Define an arbitrary domain set $\mathcal{X}$ and specify the function space of interest to be $\mathbb{R}^{\mathcal{X}} \coloneqq \{f : \mathcal{X} \to \mathbb{R}\}$. For a given subspace $V \subset \mathbb{R}^{\mathcal{X}}$, we assign to it some inner product $\langle \cdot, \cdot \rangle : V \times V \to \mathbb{R}$ that satisfies the typical properties an inner product must have for all $f, g, h \in V$ and all $a, b \in \mathbb{R}$, namely

$$\begin{aligned} \langle f, g \rangle &= \langle g, f \rangle && \text{(Symmetry)} \\ \langle af + bg, h \rangle &= a\langle f, h \rangle + b\langle g, h \rangle && \text{(Linearity)} \\ \langle f, f \rangle &> 0 && \text{(Positive-Definiteness)} \end{aligned}$$

where the Positive-Definiteness property must hold for for all $f \in V \setminus \{f_0\}$ where $f_0(x) \coloneqq 0$ for all $x \in \mathcal{X}$. Using this inner product, we define the norm of an element $f \in V$ to be $\|f\| \coloneqq \sqrt{\langle f, f \rangle}$. The tuple $(\mathcal{X}, V, \langle \cdot, \cdot \rangle)$ give us a normed vector space, which is itself a metric space for the distance measure induced by the norm. If this normed vector space $(\mathcal{X}, V, \langle \cdot, \cdot \rangle)$ is complete with respect to its norm, we call this a Hilbert space, which can also be viewed as a special case of a Banach space.

We define the evaluation functional $E_x : V \to \mathbb{R}$ to be such that $E_x[f] \coloneqq f(x)$ for some point $x \in \mathcal{X}$ and for all $f \in V$. Notice that this functional is indeed linear because if we define $f(x) = ag(x) + bh(x)$ for some other $g, h \in V$ and $a, b \in \mathbb{R}$, then $E_x[f] = f(x) = ag(x) + bh(x) = aE_x[g] + bE_x[h]$. The evaluation functional is bounded if there exists an $M_x > 0$ for a given x such that

$$|E_x[f]| \leq M_x \|f\|$$

We impose the requirement that our Hilbert space $(\mathcal{X}, V, \langle \cdot, \cdot \rangle)$ has a bounded evaluation functional, allowing us to use Theorem 2.1 to construct a *reproducing* property for the resulting space [4, 5].

Theorem 2.1: Riesz Representation Theorem for Hilbert Spaces

For a Hilbert space $(\mathcal{X}, V, \langle \cdot, \cdot \rangle)$ and any bounded linear functional $\phi : V \to \mathbb{R}$, $\exists g \in V$ such that $\forall f \in V$

$$\phi[f] = \langle f, g \rangle$$

Using Theorem 2.1 and our evaluation functional, we observe that for a given x, $\exists e_x \in V$ such that $E_x[f] = f(x) = \langle f, e_x \rangle \ \forall f \in V$. We define a kernel $k : V \times V \to \mathbb{R}$ as $k(x, y) = E_x[e_y] = e_y(x) = \langle e_y, e_x \rangle$ for all $x, y \in \mathcal{X}$. Notice that this kernel is symmetric and positive definite by the properties of the inner product, meaning that $E_x[e_y] = E_y[e_x]$. Further notice that $\forall f \in V$, $\langle f, k(x, \cdot) \rangle = \langle f, e_x \rangle = f(x)$, making it clear our kernel $k(\cdot, \cdot)$ has the ability to reproduce a given function evaluated at some point. The resulting Hilbert space, with the bounded evaluation functional assumption and the resulting reproducing kernel $k(\cdot, \cdot)$ is a Reproducing Kernel Hilbert Space (RKHS) and can be encapsulated with the tuple $(\mathcal{X}, V, \langle \cdot, \cdot \rangle, k)$ [6, 7].

3 Expectation Approximations using RKHS

Expectations over Marginals

Let us assume we are given some set E with its corresponding *σ-algebra* $\mathcal{E}$. We define $\mathcal{P}$ as the set of marginal probability distributions with respect to $\mathcal{E}$ on E. Choose some $p \in \mathcal{P}$ and define the *probability space* $(\mathcal{E}, E, p)$, where p is viewed as a *probability measure*. The expectation of some function within this probability space is equivalent to

$$\mathbb{E}_{x\sim p}[f(x)] \coloneqq \int_E f(x)dp(x)$$

Let us define a separate measure space $(\mathcal{E}, E, \nu)$ where ν need not be a probability measure but $\exists M_\nu > 0$ such that $\nu(E) \leq M_\nu$. Define V as the set of functions $f : E \to \mathbb{R}$ that are integrable with respect to the meaures p and ν within E with the requirement that $\int_E f dp < \infty$ and $\int_E f d\nu < \infty$. Define an inner product $\langle \cdot, \cdot \rangle_\nu$ to be

$$\langle f, g \rangle_\nu \coloneqq \int_E f(x)g(x)d\nu(x)$$

for all $f, g \in V$, allowing us to define the norm induced by the inner product as $\|f\|_\nu \coloneqq \sqrt{\langle f, f \rangle_\nu}$. Notice that our function space V ensures that $\langle f, g \rangle_\nu < \infty$. From our discussion earlier on RKHS and the boundedness of $\langle \cdot, \cdot \rangle_\nu$, we know there exists some bounded kernel k_ν on the set E such that $\forall f \in V$ and a given $x \in E$, $f(x) = \langle f, k_\nu(x, \cdot) \rangle$. We construct a RKHS represented by the tuple $(E, V, \langle \cdot, \cdot \rangle_\nu, k_\nu)$. Let us now define the true and empirical *mean maps* [2, 8, 1] as follows

$$\mu_x(s) \coloneqq \mathbb{E}_{x\sim p}[k_\nu(x, s)] \qquad \text{(True Mean Map)}$$

$$\hat{\mu}_x(s) \coloneqq \frac{1}{m}\sum_{i=1}^{m} k_\nu(x_i, s) \qquad \text{(Empirical Mean Map)}$$

where $D_x \coloneqq \{x_i\}_{i=1}^m$ is a dataset drawn i.i.d. from the probability distribution induced by p. One can view the expectation operation as an infinite dimensional matrix multiplication, with k_ν as the infinite dimensional matrix. The image of k_ν is V by construction, meaning that μ_x is a member of V as long as $\mathbb{E}_{x\sim p}[k_\nu(x, x)] < \infty$ holds [1]. These mean maps are convenient because they allow us to compute expectations of any function $f \in V$ by performing the inner product $\langle \mu, f \rangle_\nu$ without needing to perform the actual expectation integral, as seen in Lemma 3.1

Lemma 3.1: Computing Expectations with Mean Map

For the mean maps $\mu_x(y)$ and $\hat{\mu}_x(y)$ and any $f \in V$, performing an inner product between f and the mean maps have the following result

$$\langle \mu_x, f \rangle_\nu = \mathbb{E}_{x \sim p}[f(x)]$$
$$\langle \hat{\mu}_x, f \rangle_\nu = \frac{1}{m} \sum_{i=1}^{m} f(x_i)$$

Proof. It is relatively straight forward to show the result for the empirical mean map $\hat{\mu}_x$ and some $f \in V$.

$$\begin{aligned}
\langle \hat{\mu}_x, f \rangle_\nu &= \langle \frac{1}{m} \sum_{i=1}^{m} k_\nu(x_i, \cdot), f \rangle_\nu && \text{(Defn of empirical mean map)} \\
&= \frac{1}{m} \sum_{i=1}^{m} \langle k_\nu(x_i, \cdot), f \rangle_\nu && \text{(Linearity of Inner Product)} \\
&= \frac{1}{m} \sum_{i=1}^{m} f(x_i) && \text{(Reproducing Property)}
\end{aligned}$$

The result for the true mean map is found by expanding the expectation and moving around integrals by their linear properties.

$$\begin{aligned}
\langle \mu_x, f \rangle_\nu &= \langle \mathbb{E}_{x \sim p}[k_\nu(x, \cdot)], f \rangle_\nu && \text{(Defn of true mean map)} \\
&= \int_E f(y) \mathbb{E}_{x \sim p}[k_\nu(x, y)] d\nu(y) && \text{(Defn of } \langle \cdot, \cdot \rangle_\nu \text{)} \\
&= \int_E \int_E k_\nu(x, y) f(y) dp(x) d\nu(y) && \text{(Defn of expectation operator)} \\
&= \int_E \int_E k_\nu(x, y) f(y) d\nu(y) dp(x) && \text{(Rearrange Integrals)} \\
&= \int_E \langle f, k_\nu(x, \cdot) \rangle_\nu dp(x) && \text{(Defn of } \langle \cdot, \cdot \rangle_\nu \text{)} \\
&= \int_E f(x) dp(x) && \text{(Reproducing Property)} \\
&= \mathbb{E}_{x \sim p}[f(x)] && \text{(Defn of Expectation under } p \text{)}
\end{aligned}$$

■

The result in Lemma 3.1 give us an interesting tool for computing expectations using the true and empirical mean maps. Indeed, these mean maps have a variety of appealing benefits as mentioned in [1], one of them being that it is possible to choose a kernel k_ν such that we guarantee specific distributions map

to distinct functions in an RKHS via the mean map μ_x, making μ_x an injective map from $\mathcal{P} \to V$. We call these appropriate choices for k_ν *characteristics*. This property holds for a variety of common kernels on $\mathbb{R}^d$, the Gaussian, Laplace, and B-spline kernels being examples [8]. In the context where we do not know our true probability distribution p, making use of the empirical mean map $\hat{\mu}_x$ is sufficient by Lemma 3.2 for a sufficiently large empirical dataset.

Lemma 3.2: Empirical Mean Map convergence to True Mean Map

The empirical mean map $\hat{\mu}_x(y)$ converges to the true mean map $\mu_x(y)$ in the RKHS norm $\|\cdot\|_\nu$ with a rate of $O_p(m^{-\frac{1}{2}})$ for an empirical dataset size m.

Proof. Recall that k_ν is non-zero and bounded above within E, meaning that $\exists M > 0$ such that $k_\nu(x, y) \leq M$ for all $x, y \in E$. Also notice that for each $x_i \in D_x$ and some $s \in E$, $\mathbb{E}[k_\nu(x_i, s)] = \mathbb{E}_{x\sim p}[k_\nu(x, s)] = \mu_x(s)$ since D_x is comprised of i.i.d. samples drawn from p. By the boundedness property of k_ν on E, we know that $K(x_i, x) \in [0, M]$ for all $x \in E$ and any $x_i \in D_x$. Fix $x \in E$ and see that by Hoeffding's inequality, we have that

$$\mathbb{P}\left(|\hat{\mu}_x(x) - \mu_x(x)| \geq \epsilon\right) \leq 2\exp\left(-\frac{2m\epsilon^2}{M^2}\right)$$

for some $\epsilon > 0$. We then bound the probability that the absolute difference between $\hat{\mu}_x$ and μ_x is less than some ϵ by recalling that $\mathbb{P}(e) = 1 - \mathbb{P}(e^c)$ for some event e, thus resulting in

$$\begin{aligned}\mathbb{P}\left(|\hat{\mu}_x(x) - \mu_x(x)| \leq \epsilon\right) &= 1 - \mathbb{P}\left(|\hat{\mu}_x(x) - \mu_x(s)| \geq \epsilon\right) \\ &\geq 1 - 2\exp\left(-\frac{2m\epsilon^2}{M^2}\right)\end{aligned}$$

For some $0 < \delta < 1$, assume with probability at least $1 - \delta$ that $|\hat{\mu}_x(x) - \mu_x(x)| \leq \epsilon$ holds. Using the bound above, we choose $1 - 2\exp\left(-\frac{2m\epsilon^2}{M^2}\right) = 1 - \delta$, resulting in $\epsilon = M\sqrt{\frac{1}{2m}\ln\left(\frac{2}{\delta}\right)}$. Thus we have that for any $x \in E$ that

$$|\hat{\mu}_x(x) - \mu_x(x)| \leq M\sqrt{\frac{1}{2m}\ln\left(\frac{2}{\delta}\right)}$$

Define $t^* = \arg\sup_{t\in E} |\hat{\mu}_x(t) - \mu_x(t)|$ and notice that the bound above still holds for $|\hat{\mu}_x(t^*) - \mu_x(t^*)|$ since t^* is an element of E. Using the definition for $\|\cdot\|_\nu$, we then show that

$$\begin{aligned}\|\mu_x - \hat{\mu}_x\|_\nu &\leq \|\|\hat{\mu}_x(t^*) - \mu_x(t^*)|\mathbf{1}\|_\nu \\ &= |\hat{\mu}_x(t^*) - \mu_x(t^*)| \, \|\mathbf{1}\|_\nu \\ &= |\hat{\mu}_x(t^*) - \mu_x(t^*)|\nu(E)^{\frac{1}{2}} \\ &\leq M_\nu^{\frac{1}{2}} M \sqrt{\frac{1}{2m} \ln\left(\frac{2}{\delta}\right)}\end{aligned}$$

where $\mathbf{1}(y) = 1$ for all $y \in E$. We see then that $\|\mu_x - \hat{\mu}_x\|_\nu \in O\left(m^{-\frac{1}{2}}\right)$ for a fixed δ. Thus, in probability we find that the empirical mean map $\hat{\mu}_x$ converges to the true mean map μ_x with a rate of $O_p\left(m^{-\frac{1}{2}}\right)$ with respect to the RKHS norm $\|\cdot\|_\nu$. ■

Higher Order Moments and Joint Distributions

Let us consider two reproducing kernel Hilbert spaces $(E_X, V_X, \langle\cdot,\cdot\rangle_X, k_X)$ and $(E_Y, V_Y, \langle\cdot,\cdot\rangle_Y, k_Y)$ respectively dependent on the measure spaces $(\mathcal{E}_X, E_X, \nu_X)$ and $(\mathcal{E}_Y, E_Y, \nu_Y)$ with the property that the measures are bounded on their respective spaces. Suppose we have a joint distribution p_{XY} and the marginal distribution p_X. We define the *uncentered covariance operator* $C_{XX}(s,t) \coloneqq \mathbb{E}_{x\sim p_X}\left[k_X(x,s)k_X(x,t)\right]$ for all $s,t \in E_X$ and find that this construction results in the property seen in Lemma 3.3.

Lemma 3.3: Inner Product with Covariance Operator

For some $f \in V_X$, the uncentered covariance operator C_{XX} has the property that

$$\langle f, C_{XX} f\rangle_X = \mathbb{E}_{x\sim p_X}[(f(x))^2]$$

Proof. Let us first recognize that the operation $C_{XX}f$ is equivalent to an integral operator in the measure space related to ν_X, thus giving us that $C_{XX}f \coloneqq \int_{E_X} C_{XX}(\cdot,s)f(s)d\nu_X(s)$. Using this definition and the definition of $\langle\cdot,\cdot\rangle_X$ based on measure ν_X, we work out the following

$$
\begin{aligned}
\langle f, C_{XX} f \rangle_X &= \int_{E_X} f(y) \left(C_{XX} f\right)(y) d\nu_X(y) && \text{(Defn } \langle \cdot, \cdot \rangle_X\text{)} \\
&= \int_{E_X} f(y) \int_{E_X} C_{XX}(y, s) f(s) d\nu_X(s) d\nu_X(y) && \text{(Defn } C_{XX} f\text{)} \\
&= \int_{E_X} f(y) \int_{E_X} f(s) \int_{E_X} k_X(x, y) k_X(x, s) dp_X(x) d\nu_X(s) d\nu_X(y) \\
& && \text{(Defn } C_{XX}\text{)} \\
&= \int_{E_X} \left(\int_{E_X} k_X(x, s) f(s) d\nu_X(s) \right)^2 dp_X(x) \\
& && \text{(Integral rearranging)} \\
&= \int_{E_X} f(x)^2 dp_X(x) && \text{(Reproducing property)} \\
&= \mathbb{E}_{x \sim p_X} [(f(x))^2]
\end{aligned}
$$

■

We now define the *cross-covariance* operator $C_{XY}(s, t) \coloneqq \mathbb{E}_{(x,y) \sim p_{XY}} [k_X(x, s) k_Y(y, t)]$. Using this definition, a similar property to that seen in Lemma 3.3 is shown below in Lemma 3.4.

Lemma 3.4: Inner Product with Covariance Operator

For some $f \in V_X$ and $g \in V_Y$, the uncentered cross-covariance operator C_{XY} has the property that

$$\langle f, C_{XY} g \rangle_X = \mathbb{E}_{(x,y) \sim p_{XY}} [f(x) g(y)]$$

Proof. Similar to Lemma 3.3, the operation $C_{XY} g$ is equivalent to an integral operator in the measure space related to ν_Y, thus allowing the definition $C_{XY} g \coloneqq \int_{E_Y} C_{XY}(\cdot, s) g(s) d\nu_Y(s)$. Using this definition and the definition of $\langle \cdot, \cdot \rangle_X$ based on measure ν_X, we can work out the following

$$
\begin{aligned}
\langle f, C_{XY} g \rangle_X &= \int_{E_X} f(t) \left(C_{XY} g\right)(t) d\nu_X(t) && \text{(Defn } \langle \cdot, \cdot \rangle_X\text{)} \\
&= \int_{E_X} f(t) \int_{E_Y} C_{XY}(t,s) g(s) d\nu_Y(s) d\nu_X(t) && \text{(Defn } C_{XY} g\text{)} \\
&= \int_{E_X} f(t) \int_{E_Y} g(s) \int_{E_Y \times E_X} k_X(x,t) k_Y(y,s) dp_{XY}(x,y) d\nu_Y(s) d\nu_X(t) \\
& && \text{(Defn } C_{XY}\text{)} \\
&= \int_{E_Y \times E_X} \left(\int_{E_X} k_X(x,t) f(t) d\nu_X(t) \right) \left(\int_{E_Y} k_Y(y,s) g(s) d\nu_Y(s) \right) dp_{XY}(x,y) \\
& && \text{(Integral rearranging)} \\
&= \int_{E_Y \times E_X} f(x) g(y) dp_{XY}(x,y) && \text{(Reproducing property)} \\
&= \mathbb{E}_{(x,y) \sim p_{XY}} [f(x) g(y)]
\end{aligned}
$$

■

Conditional Distributions

It is of interest to find some mean map $\mu_{Y|x} \in V_Y$ that is capable of representing conditional distributions $P(Y|X = x)$, for some given $x \in E_X$, such that $\langle g, \mu_{Y|x} \rangle_Y = \mathbb{E}_{Y|X=x}[g(Y)]$. It was found in [1] that we can construct the operator $\mathcal{U}_{Y|X} \coloneqq C_{YX} C_{XX}^{-1}$ and specify $\mu_{Y|x}(y) = \left(\mathcal{U}_{Y|X} k_X(x, \cdot)\right)(y)$ to obtain the desired property for the inner product $\langle g, \mu_{Y|x} \rangle_Y$. Further, it was found in [1] that using a finite dataset $D_{XY} = \{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from p_{XY}, we can estimate $\mathcal{U}_{Y|X}$ using the estimator

$$
\hat{\mathcal{U}}_{Y|X}(s,t) = \Phi(s) \left(K + \lambda m I\right)^{-1} \Upsilon^T(t) \tag{1}
$$

for $s \in E_y$ and $t \in E_x$ where $K_{ij} \coloneqq k_x(x_i, x_j)$, $\Phi(s) \coloneqq [k_Y(y_1, s), \cdots, k_Y(y_m, s)]$, $\Upsilon(t) = [k_X(x_1, t), \cdots, k_X(x_m, t)]$, and for some regularization parameter $\lambda > 0$ to help ensure the inverse exists on the finite dataset. Using this estimator, we arrive at a compact estimator for $\mu_{Y|x}$ found in Lemma 3.5.

Lemma 3.5: Empirical Conditional Mean Map Representation

Using a finite dataset $D_{XY} = \{(x_i, y_i)\}_{i=1}^m$ drawn i.i.d. from p_{XY}, a conditional mean map $\mu_{Y|x}$ has an estimate of the form

$$
\hat{\mu}_{Y|x}(y) = \sum_{i=1}^m \beta_i(x) k_Y(y_i, y)
$$

Proof. Using estimate $\hat{\mathcal{U}}_{Y|X}(s,t)$ from (1) and the definition $\mu_{Y|x}(y) = \left(\mathcal{U}_{Y|X} k_X(x,\cdot)\right)(y)$, we find $\hat{\mu}_{Y|x}(y)$ with the following steps

$$\begin{aligned}
\hat{\mu}_{Y|x}(y) &= \left(\hat{\mathcal{U}}_{Y|X} k_X(x,\cdot)\right)(y) \\
&= \int_{E_x} \hat{\mathcal{U}}_{Y|X}(y,t) k_X(x,t) d\nu_x(t) \\
&= \int_{E_x} \Phi(y) \left(K + \lambda m I\right)^{-1} \Upsilon^T(t) k_X(x,t) d\nu_x(t) \\
&= \Phi(y) \left(K + \lambda m I\right)^{-1} \underbrace{\int_{E_x} \Upsilon^T(t) k_X(x,t) d\nu_x(t)}_{\Psi^T(x)} \\
&= \Phi(y) \underbrace{\left(K + \lambda m I\right)^{-1} \Psi^T(x)}_{\boldsymbol{\beta}^T(x)} \\
&= \sum_{i=1}^{m} \beta_i(x) k_Y(y_i, y)
\end{aligned}$$

where $\boldsymbol{\beta}(x) = [\beta_1(x), \cdots, \beta_m(x)]$. ■

Lemma 3.5 is interesting since it shares a familiar appearance to the empirical mean map defined for marginals except in the case of marginals, the kernel is weighted an equal $\frac{1}{m}$ across all samples. The fact this conditional variant weights the kernel evaluations within the dataset differently makes sense since we need to capture the conditional behavior.

Now the quality of the estimate $\hat{\mathcal{U}}_{Y|X}$ is important to understand since this estimate decides how well $\hat{\mu}_{Y|x}$ approximates $\mu_{Y|x}$. From [1], we have the following convergence theorem, Theorem 3.1.

Theorem 3.1: Convergence of $\hat{\mathcal{U}}_{Y|X}$ to $\mathcal{U}_{Y|X}$

Assume $C_{YX} C_{XX}^{\frac{3}{2}}$ is Hilbert-Schmidt. Then $\left\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\right\|_{HS} \in O_p\left(\lambda^{1/2} + \lambda^{-3/2} m^{-1/2}\right)$, where $\|\cdot\|_{HS}$ is the Hilbert-Schmidt norm. If we choose the regularization term such that $\lambda \to 0$ and $m\lambda^3 \to \infty$, then $\left\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\right\|_{HS} \to 0$ in probability.

4 Value Iteration Approximation via RKHS

Within the context of reinforcement learning, we are generally first presented with a Markov Decision Process (MDP) of the form $M = (\mathcal{X}, \mathcal{A}, R, P, \gamma)$ where $\mathcal{X}$ is the set of possible states, $\mathcal{A}$ is the set of possible actions, $R : X \times A \to \mathbb{R}$ is a reward function that maps state-action pairs to some reward value, P is a

conditional distribution $P(X'|X, A)$ representing the *transition probabilities* of going from a given state-action pair $(x, a) \in \mathcal{X} \times \mathcal{A}$ to some new state $x' \in \mathcal{X}$, and $0 < \gamma < 1$ is referred to as the *discount factor*. Out ultimate goal is to find some policy $\pi : \mathcal{X} \to \mathcal{A}$ such that we obtain an optimum amount of *value*, given the properties of M, with respect to the *value function* V^π as defined below

$$V^\pi(x) \coloneqq \mathbb{E}\left[\sum_{h=0}^{\infty} \gamma^{h-1} r_h \,\middle|\, \begin{matrix} x_0 = x \\ a_h = \pi(x_h) \\ r_h = R(x_h, a_h) \end{matrix}\right] \tag{2}$$

We can reshape (2) into the form $V^\pi(x) = R(x, \pi(x)) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,\pi(x))}\left[V^\pi(x')\right]$, providing a useful recursive form. If we assume we have a policy π^* that optimizes the amount of value that can be obtained from our MDP M, this policy satisfies the Bellman Optimality equations as defined in Definition 4.1.

Definition 4.1: Bellman Optimality Equations

An optimal policy π^* has a corresponding value function $V^* \coloneqq V^{\pi^*}$ such that the following Bellman optimality conditions hold $\forall x \in \mathcal{X}$

$$\begin{aligned} V^*(x) &= \left(\mathcal{T}V^*\right)(x) \\ \pi^*(x) &= \arg\max_{a \in \mathcal{A}} \left\{R(x, a) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,a)}\left[V^*(x')\right]\right\} \end{aligned}$$

where $\mathcal{T} : \mathbb{R}^{\mathcal{X}} \to \mathbb{R}^{\mathcal{X}}$ is called the *Bellman operator* and for some $V \in \mathbb{R}^{\mathcal{X}}$ it is define as

$$\left(\mathcal{T}V\right)(x) = \max_{a \in \mathcal{A}} \left\{R(x, a) + \gamma \mathbb{E}_{x' \sim P(\cdot|x,a)}\left[V(x')\right]\right\}$$

Within the context of tackling MDPs like those defined above, there are times we do not know what the transition probabilities $P(X'|X, A)$ are. This lack of crucial information requires us either approximate the Bellman operator in some form or use other very different approaches. With the foundational work we did prior, we are able to approach approximating the Bellman operator via sampled data, using this sampled data to construct a conditional mean map that approximates the expectation term in the Bellman operator.

From here, we define two reproducing kernel Hilbert spaces $(\mathcal{X}, V_X, \langle\cdot,\cdot\rangle_X, k_X)$ and $(\mathcal{X} \times \mathcal{A}, V_{XA}, \langle\cdot,\cdot\rangle_{XA}, k_{XA})$ where $V_X \subset \mathbb{R}^{\mathcal{X}}$ and $V_{XA} \subset \mathbb{R}^{\mathcal{X}\times\mathcal{A}}$. We assume that the measure for $\langle\cdot,\cdot\rangle_X$ is ν_X and the induced norm is defined as $\|\cdot\|_X$ and similarly for the measure and norm associated with $\langle\cdot,\cdot\rangle_{XA}$. Using Lemma 3.5, we construct a dataset $D \coloneqq \{(x_i', a_i, x_i)\}_{i=1}^m$ that is drawn i.i.d. from $P(X'|X, A)$. We then define the empirical mean map to be

$$\hat{\mu}_{(x,a)}(x') \coloneqq \sum_{i=1}^{m} \beta_i(x,a) k_X(x_i', x')$$

where if we define $W \coloneqq (K + \lambda m I)^{-1}$ with $K_{ij} = k_{XA}\left((x_i, a_i), (x_j, a_j)\right)$ and regularization parameter λ, we define $\beta_i(x,a)$ to be

$$\begin{aligned}
\beta_i(x,a) &\coloneqq \sum_{j=1}^{m} W_{ij} \int_{\mathcal{X}\times\mathcal{A}} k_{XA}\left((x_j,a_j),(s,t)\right) k_{XA}\left((x,a),(s,t)\right) d\nu_{XA}(s,t) \\
&= \sum_{j=1}^{m} W_{ij} k_{XA}\left((x_j,a_j),(x,a)\right) \qquad \text{(Reproducing property)}
\end{aligned}$$

By Theorem 3.1, we know that $\hat{\mu}_{(x,a)}$ will converge to $\mu_{(x,a)}$ for all $(x,a) \in \mathcal{X} \times \mathcal{A}$ as we increase our dataset size m. Let us now define $\hat{\mathcal{E}}_{(x,a)}[f] \coloneqq \langle \hat{\mu}_{(x,a)}, f \rangle_X$. When $f \in V_X$, we know that $\hat{\mathcal{E}}_{(x,a)}[f] \approx \mathbb{E}_{x' \sim P(\cdot|x,a)}\left[f(x')\right]$. When $f \notin V_X$, it is clear the accuracy of this operator is dependent on how close f is to some element in V_X. Regardless, we are now able to replace $\mathbb{E}_{x' \sim P(\cdot|x,a)}\left[V(x')\right]$ with $\hat{\mathcal{E}}_{(x,a)}[V]$ in the Bellman operator, producing an approximate Bellman operator defined as

$$\left(\hat{\mathcal{T}}V\right)(x,a) \coloneqq R(x,a) + \gamma \hat{\mathcal{E}}_{(x,a)}[V] \tag{3}$$

The benefit of this new expectation approximation is that after constructing the mean maps, each expectation has a cost of $O(m)$. Now work in [2] provides holistic bound for the value function approximation and resulting greedy policy with respect to the true optimal policy π^* and true optimal value function V^*. Let us first define a value iteration algorithm based on the approximate Bellman operator. Assume first that we are given an arbitrary initial value function $\hat{V}_0$. Using the approximate Bellman operator $\hat{\mathcal{T}}$, we update our value function using the recursion $\hat{V}_{j+1} \leftarrow \hat{\mathcal{T}}\hat{V}_j$ where $\hat{V}_j$ is the j^{th} estimate. Work in [2] found that our approximate Bellman operator $\hat{\mathcal{T}}$ is a γ-contraction in the infinity norm for bounded Banach functions. This implies that $\left\|\hat{\mathcal{T}}V - \hat{\mathcal{T}}V'\right\| \leq \gamma \left\|V - V'\right\|$ for any V, V' that are within a Banach space of bounded functions.

Using our value iteration algorithm, it is possible to show that with the γ-contraction property of $\hat{\mathcal{T}}$ that we converge to a fixed point $\hat{V}^*$. In [2], they bound the difference between the j^{th} iterate $\hat{V}_j$ and the fixed point using

$$\left\|\hat{V}_j - \hat{V}^*\right\|_\infty \leq \frac{\gamma^j}{1-\gamma} \left\|\hat{V}_1 - \hat{V}_0\right\|_\infty$$

clearly showing the distance in the infinity norm between our iterate $\hat{V}_j$ and the fixed point $\hat{V}^*$ shrink with increasing number of iterations. Now for any

estimate $\hat{V}$, define $\widetilde{V} \coloneqq \Pi_X V$ as the projection of $\hat{V}$ into V_X. Now suppose we perform the value iteration for κ steps, arriving at our estimate for the optimal value as $\hat{V}_\kappa$. Define $\hat{\pi}_\kappa(x) \coloneqq \arg\max_{a \in \mathcal{A}} R(x,a) + \gamma \hat{\mathcal{E}}_{(x,a)}[\hat{V}_\kappa]$ as the greedy policy associated with $\hat{V}_\kappa$. From [2], we have Theorem 4.1.

Theorem 4.1: Error of Value Iteration with Approximate Bellman Operator

An optimal policy π^* has a corresponding value function $V^* \coloneqq V^{\pi^*}$ such that the following Bellman optimality conditions hold $\forall x \in \mathcal{X}$

$$\|V^{\pi_\kappa} - V^*\|_\infty \leq \frac{2\gamma}{(1-\gamma)^2} \left(\gamma^\kappa \left\| \hat{V}_1 - \hat{V}_0 \right\|_\infty + 2 \left\| V^* - \widetilde{V}^* \right\|_\infty + \sup_{(x,a)} \left\| \hat{\mu}_{(x,a)} - \mu_{(x,a)} \right\|_X \left\| \widetilde{V}^* \right\|_X \right)$$

Theorem 4.1 provides a holistic view of where error might arise. The first term expresses the initial error generated between the initial guess value function $\hat{V}_0$ and the first one generated using the approximate Bellman operator, $\hat{V}_1$. The second term corresponds to the error between the optimal solution V^* and the projection $\Pi_X V^*$ into the set V_X. This second term can be reduced by crafting a richer function RKHS V_X. The latter term corresponds to the error with our empirical conditional mean map. As we increase the dataset size we use to construct it and choose a regularization parameter λ accordingly, this term should reduce to 0 in probability.

References

[1] L. Song, J. Huang, A. Smola, and K. Fukumizu, "Hilbert space embeddings of conditional distributions with applications to dynamical systems," 2009. [Online]. Available: http://jonathan-huang.org/research/pubs/icml09/icml09.pdf

[2] S. Grunewalder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton, "Modelling transition dynamics in mdps with rkhs embeddings," 2012. [Online]. Available: https://arxiv.org/pdf/1206.4655.pdf

[3] "Mdp preliminaries," 2019. [Online]. Available: https://nanjiang.cs.illinois.edu/files/cs598/note1.pdf

[4] "The riesz representation theorem for hilbert spaces." [Online]. Available: http://mathonline.wikidot.com/the-riesz-representation-theorem-for-hilbert-spaces

[5] "Linear functionals and bounded linear functionals." [Online]. Available: http://mathonline.wikidot.com/linear-functionals-and-bounded-linear-functionals

[6] "Reproducing kernel hilbert spaces," 2006. [Online]. Available: http://www.mit.edu/~9.520/spring06/Classes/class03.pdf

[7] A. Gretton, "Introduction to rkhs, and some simple kernel algorithms," 2019. [Online]. Available: http://www.gatsby.ucl.ac.uk/~gretton/coursefiles/lecture4_introToRKHS.pdf

[8] L. Song, A. Gretton, and C. Guistrin, "Nonparametric tree graphical models via kernel embeddings," 2010. [Online]. Available: http://proceedings.mlr.press/v9/song10a/song10a.pdf

[9] C. S. Kubrusly, *Measure Theory, A First Course*. Elsevier, Inc., 2007.

[10] D. P. Bertsekas and J. N. Tsitsiklis, *Introduction to Probability*. Athena Scientific, 2008.

[11] "Concentration inequalities and multi-armed bandits," 2019. [Online]. Available: https://nanjiang.cs.illinois.edu/files/cs598/note_bandit.pdf

[12] "Banach space." [Online]. Available: http://mathworld.wolfram.com/BanachSpace.html

[13] "Convergence in probability." [Online]. Available: https://en.wikipedia.org/wiki/Convergence_of_random_variables#Convergence_in_probability

[14] "Banach fixed-point theorem." [Online]. Available: https://en.wikipedia.org/wiki/Banach_fixed-point_theorem