Modelling transition dynamics in MDPs with RKHS embeddings

Steffen Grünewälder[1]   STEFFEN@CS.UCL.AC.UK
Guy Lever[1]   G.LEVER@CS.UCL.AC.UK
Luca Baldassarre   L.BALDASSARRE@CS.UCL.AC.UK
Massimilano Pontil   M.PONTIL@CS.UCL.AC.UK
Arthur Gretton*   ARTHUR.GRETTON@GMAIL.COM

CSML and *Gatsby Unit, University College London, UK, * MPI for Intelligent Systems

Abstract

We propose a new, nonparametric approach to learning and representing transition dynamics in Markov decision processes (MDPs), which can be combined easily with dynamic programming methods for policy optimisation and value estimation. This approach makes use of a recently developed representation of conditional distributions as *embeddings* in a reproducing kernel Hilbert space (RKHS). Such representations bypass the need for estimating transition probabilities or densities, and apply to any domain on which kernels can be defined. This avoids the need to calculate intractable integrals, since expectations are represented as RKHS inner products whose computation has linear complexity in the number of points used to represent the embedding. We provide guarantees for the proposed applications in MDPs: in the context of a value iteration algorithm, we prove convergence to either the optimal policy, or to the closest projection of the optimal policy in our model class (an RKHS), under reasonable assumptions. In experiments, we investigate a learning task in a typical classical control setting (the under-actuated pendulum), and on a navigation problem where only images from a sensor are observed. For policy optimisation we compare with least-squares policy iteration where a Gaussian process is used for value function estimation. For value estimation we also compare to the NPDP method. Our approach achieves better performance in all experiments.

1. Introduction

1.1. Preliminaries

Throughout we denote expectations by $\mathbb{E}[\cdot]$, and the probability over events by $\mathbb{P}(\cdot)$. We denote by $\mathcal{B}(\mathcal{X})$ and $\mathcal{C}_b(\mathcal{X})$ the Banach spaces of bounded functions and bounded continuous functions on $\mathcal{X}$, each equipped with the sup-norm $||\cdot||_\infty$.

We consider in particular the problem in which we control a trajectory $\{x_t\}_{t=0}^\infty$ over $\mathcal{X}$ by sequentially choosing actions $a_t \in \mathcal{A}$ at each time step $t \geq 0$, once x_t is revealed, after which we receive a reward $r_{t+1} = r(x_t, a_t)$. We denote a set of deterministic *policies* $\Pi = \mathcal{A}^\mathcal{X}$. The objective is to find a policy π which maximises the expected sum of rewards obtained by following π: $\mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_{t+1}(X_t, A_t) | X_0 = x, A_t = \pi(X_t)\right]$.

For a policy $\pi \in \Pi$ we denote the associated *value function*,

$$V^\pi(x) := \mathbb{E}\left[\sum_{t=0}^\infty \gamma^t r_{t+1}(X_t, A_t) | X_0 = x, A_t = \pi(X_t)\right],$$

and recall that $V^\pi(x) = r(x, \pi(x)) + \gamma \mathbb{E}_{X \sim P(\cdot|x,\pi(x))}[V^\pi(X)]$. We define the optimal value function $V^*(x) := \max_{\pi \in \Pi} V^\pi(x)$ for all $x \in \mathcal{X}$, and an optimal policy to be any π^* such that $\pi^* \in \operatorname{argmax}_{\pi \in \Pi} V^\pi(x)$ for all $x \in \mathcal{X}$. For a given *action-value function* $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ we define the *greedy policy* w.r.t. Q by $\pi_Q(x) := \operatorname{argmax}_{a \in \mathcal{A}} Q(x, a)$ (choosing arbitrarily in the case of a tie) and the optimal action-value function,

$$Q^*(x, a) := r(x, a) + \gamma \mathbb{E}_{X \sim P(\cdot|x,a)}[V^*(X)], \quad (1)$$

so that $\pi^* = \pi_{Q^*}$ (see e.g. (Szepesvari, 2009) for this background). We require the following well-known result, which is proved in the Appendix for reference (Grünewälder et al., 2012):

---

[1]**Equal contribution.**

Lemma 1.1. *(Singh & Yee, 1994)[Corollary 2] For any action-value function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$, the greedy policy π_Q satisfies $||V^{\pi_Q} - V^*||_\infty \leq \frac{2}{1-\gamma}||Q^* - Q||_\infty$.*

We are interested in the case where P is unknown but a sample $\mathcal{S} := \{(x_i, a_i, x'_i)\}_{i=1}^m$ is provided, drawn i.i.d. from a distribution $\widetilde{P}$ such that $\widetilde{P}(X'_i = x'_i|X_i = x_i, A_i = a_i) = P(x'_i|x_i, a_i)$ for all i (the marginal probabilities need not match). Note the abuse of notation here – subscripts index samples and not time steps.

1.2. Overview of the approach

A number of recent studies have focused on efficient evaluation of conditional expectations on functions that are "well behaved" in the sense that they belong to a reproducing kernel Hilbert space (RKHS). These approaches have been particularly successful in performing inference in graphical models, where the model parameters are learned nonparametrically from data (Song et al., 2010b; 2009; 2011). The key insight in these works is that conditional probabilities can be represented as functions in an RKHS, called *conditional distribution embeddings*. The conditional expectation of any function in the RKHS then becomes a linear operation, where we take the inner product with the appropriate distribution embedding.

Many methods for solving problems in MDPs require the computation of expectations of functions (value functions for example) with respect to transition dynamics, and so (approximations of) the operators

$$f \mapsto \mathbb{E}_{X \sim P(\cdot|x,a)}[f(X)] \tag{2}$$

are required. A direct but computationally costly approach would be to first learn a conditional density estimate (difficult in high dimensions), followed by (possibly intractable) integrals to compute the expectation. By contrast, our approach is a two stage process for learning in MDPs: we first use the theory of RKHS embeddings to estimate the operators (2) *directly* (over a specific class of functions in an RKHS), then use these estimated operators in standard approaches for solving MDPs – here we consider dynamic programming methods for value estimation and policy optimisation. The application to dynamic programming is described in more detail in Sec. 3.

1.3. Advantages of the approach

A direct kernel-based approach has a number of advantages. First, like density estimates, conditional embeddings can be learned from a training sample: we do not need to address the problem of modeling system dynamics, such as the differential equations governing

a robot arm. Unlike density estimates, however, distribution embedding estimates do not scale poorly with the dimension d of the underlying space: the risk of a kernel density estimate increases as $O(m^{-4/(4+d)})$ when the optimal bandwidth is used (Wasserman, 2006)[Sec. 6.5]. By contrast, the rate of convergence for conditional mean embeddings is independent of the dimension of the underlying space (Song et al., 2010b)[Thm. 1].

Second, the solution to many control problems involves computation of high dimensional integrals to obtain expectations, which is prohibitively costly. By contrast, RKHS embeddings explicitly provide a representation of the expectation operator as an RKHS inner product, which reduces calculating expectations to a computation of *linear complexity* in the number of training points used to represent the embedding, and avoids any intermediate problems such as density estimation and sample selection for numerical integration. Thus, the approach provides a framework for alleviating the curse of dimensionality in MDPs (particularly if, for example, sparsification of the embedding is considered, which we address briefly in the Appendix (Grünewälder et al., 2012)). The conditional distribution embeddings themselves may be computed exactly at cost cubic in the training sample size, and approximated to good accuracy at linear cost.

A third advantage is that we can provide convergence results in the infinite sample case. Thm. 3.2 demonstrates how a performance guarantee for value iteration using embeddings decomposes into guarantees for value iteration and gurantees for the embeddings, upper bounding the difference $||V^{\widehat{\pi}_\kappa} - V^*||_\infty$ between the optimal value V^* and the value $V^{\widehat{\pi}_\kappa}$ of the policy $\widehat{\pi}_\kappa$ found by performing value iteration using the embeddings after κ iterations. This bound contains a term involving how well we can approximate V^* in our model class (a chosen RKHS) – which usually corresponds to smoothness assumptions on V^* – and can decrease by increasing the richness of the RKHS. A second term captures how quickly we can learn the embeddings for the operator (2) over functions in the chosen RKHS. This bound can be specialised to give convergence guarantees for specific settings by plugging in guarantees for the two components: in Corollary 3.3, we specialise to the common setting of finite state space and positive definite kernel and obtain that $||V^{\widehat{\pi}_\kappa} - V^*||_\infty \to 0$.

As a final advantage, the method applies wherever kernels may be defined, including on high dimensional or continuous state spaces, manifolds (kernels on the surface of a sphere (Wendland, 2005) are of particu-

lar interest in robotics), and partially observable tasks where only sensor measurements are available.

1.4. Relation to existing methods

Kernel methods have become increasingly popular in RL. Methods include kernel LSTD (Xu et al., 2005) and GPTD (Engel et al., 2005). Both can be used to estimate (q-)values and they differ in this mainly through the regulariser (Taylor & Parr, 2009). Based on the q-value estimates it is possible to optimise the policy. Other related approaches include Rasmussen & Kuss (2003), Deisenroth et al. (2009), Ormoneit & Sen (1999) and Kroemer & Peters (2011). Here, transition models (densities) are learned with the help of Gaussian processes or kernel density estimates. Using them for value estimation or policy optimisation usually leads to difficult integration to be solved numerically via, e.g., an intermediate sampling method. In contrast we use kernels to directly learn the expectation operators and avoid numerical integration. Finally, in (Parr et al., 2008) a way is proposed to approximate expectations in a low dimensional state representation. In contrast to our approach the paper assumes that the true expectation is known.

2. RKHS embeddings of transition probability kernels

Given a set $\mathcal{Z}$ and a positive semi-definite (p.s.d.) kernel $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ (see e.g. Steinwart & Christmann, 2008, for details) we denote by $\mathcal{H}_K \subseteq \mathbb{R}^{\mathcal{Z}}$ its unique reproducing kernel Hilbert space (RHKS), and by $\langle \cdot, \cdot \rangle_K$ the inner product in $\mathcal{H}_K$. Due to the reproducing property of K in $\mathcal{H}_K$ we have $h(x) = \langle K(x, \cdot), h \rangle_K$ for all $h \in \mathcal{H}_K$. We recall the notion of a universal kernel: given a Banach space of functions $\mathcal{F} \subseteq \mathbb{R}^{\mathcal{Z}}$ a kernel is $\mathcal{F}$-*universal* if $\mathcal{H}_K$ is dense in $\mathcal{F}$. We denote $\rho_K := \sup_{z \in \mathcal{Z}} \sqrt{K(z,z)}$ and refer to kernels K such that $\rho_K < \infty$ as bounded kernels.

Following Sriperumbudur et al. (2010), given any probability distribution P and p.s.d. kernel K on a set $\mathcal{Z}$ a *distribution embedding* of P in $\mathcal{H}_K$ is an element $\mu \in \mathcal{H}_K$ such that $\langle \mu, h \rangle_K = \mathbb{E}_{Z \sim P}[h(Z)]$ for all $h \in \mathcal{H}_K$. In our application, given p.s.d. kernels $L : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and $K : (\mathcal{X} \times \mathcal{A}) \times (\mathcal{X} \times \mathcal{A}) \to \mathbb{R}$, we are interested in the embedding of the expectation operator (2) corresponding to the state transition probability kernel P, over the domain $\mathcal{H}_L$; that is, an element $\mu_{(x,a)} \in \mathcal{H}_L$ such that $\langle \mu_{(x,a)}, f \rangle_L = \mathbb{E}[f(X_{t+1})|X_t = x, A_t = a]$, for all $f \in \mathcal{H}_L$ and for all $t \geq 0$ – recall that the Markov property implies such a $\mu_{(x,a)}$ is independent of time. Recalling Sec. 1.1, given the sample $\mathcal{S}$, we will consider a sample-based estimate of the expecta-

tion operator (2). This will be achieved by identifying an element $\overline{\mu}_{(x,a)} \in \mathcal{H}_L$ such that, for all $f \in \mathcal{H}_L$, $\langle \overline{\mu}_{(x,a)}, f \rangle_L$ approximates $\mathbb{E}_{X \sim P(\cdot|x,a)}[f(X)]$. Following (Song et al., 2009; 2010b) an estimate is

$$\overline{\mu}_{(x,a)} := \sum_{i=1}^m \alpha_i(x,a) L(x_i', \cdot) \in \mathcal{H}_L, \qquad (3)$$

where $\alpha_i(x,a) = \sum_{j=1}^m W_{ij} K((x_j, a_j), (x,a))$, and where $\boldsymbol{W} := (\boldsymbol{K} + \lambda m \boldsymbol{I})^{-1}$, $\boldsymbol{K} = (K((x_i, a_i), (x_j, a_j)))_{ij=1}^m$, and λ is a regularization parameter. We assume w.l.o.g. $x_i' \neq x_j'$ for all x_i', x_j' in the expansion (3).[1] In some situations, the estimate (3) is consistent in the RHKS norm sense and uniformly over $\mathcal{X} \times \mathcal{A}$: the following result, proved in the appendix, follows directly from (Song et al., 2010b)[Thm. 1].

Lemma 2.1. *Suppose K is a bounded kernel and the conditions of (Song et al., 2010b)[Thm. 1] are satisfied.*[2] *Then $\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \{ ||\mu_{(x,a)} - \overline{\mu}_{(x,a)}||_L \} \in \mathcal{O}_{\widetilde{P}}(\lambda^{\frac{1}{2}} + \lambda^{-\frac{3}{2}} m^{-\frac{1}{2}})$, and thus by choosing $\lambda \to 0$, $\lambda^3 m \to \infty$ we have that, for any $\epsilon > 0$,*

$$\mathbb{P}_{\mathcal{S} \sim \widetilde{P}^m} \left(\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} ||\mu_{(x,a)} - \overline{\mu}_{(x,a)}||_L > \epsilon \right) \to 0.$$

By the reproducing property of L, we have

$$\langle \overline{\mu}_{(x,a)}, f \rangle_L = \sum_{i=1}^m \alpha_i(x,a) f(x_i')$$

In this work, for theoretical analysis, we consider a normalised version of (3):

$$\widehat{\mu}_{(x,a)} := \sum_{i=1}^m \widehat{\alpha}_i(x,a) L(x_i', \cdot) \in \mathcal{H}_L, \qquad (4)$$

where $\widehat{\alpha}_i(x,a) = \frac{\alpha_i(x,a)}{\sum_{j=1}^m |\alpha_j(x,a)|}$. This is a technical consideration which will later ensure that we can define

---

[1] We can otherwise form a new expansion in which the x_i' are unique by summing any $\alpha_i(x,a)$ as necessary.

[2] These conditions require that the mapping $(x,a) \mapsto \mathbb{E}_{X \sim P(\cdot|(x,a))}[f(X)]$ be an element of $\mathcal{H}_K$ for all $f \in \mathcal{H}_L$, and that the operator $C_{YX} C_{XX}^{-3/2}$ be Hilbert-Schmidt, where C_{YX} and C_{XX} are covariance operators: see (Song et al., 2009) or Appendix D.1 for details (Grünewälder et al., 2012). The first condition is a smoothness assumption on the distribution, and for the convergence guarantee of Corollary 3.3 we specialise to the simple setting of finite state space, in which case this condition is trivially satisfied. The second condition is guaranteed in our case when, for example, the marginal density of the initial state X from $\widetilde{P}$ is bounded away from zero and the RKHSs $\mathcal{H}_K$, $\mathcal{H}_L$ are of finite dimensionality.

Algorithm 1 Estimate Conditional Expectation

input Sample of transitions $\mathcal{S} := \{(x_1, a_1, x_1'), \ldots, (x_m, a_m, x_m')\}$, kernel K on $\mathcal{X} \times \mathcal{A}$ and a kernel L on states $\mathcal{X}$
output A conditional expectation estimate $\overline{\mu}_{(x,a)}$
   Build kernel matrix $\boldsymbol{K}$ for samples $\{(x_1, a_1), \ldots, (x_m, a_m)\}$
   Calculate coefficient vector $\alpha_i := \sum_{j \le m} \boldsymbol{W}_{ij} K((x_j, a_j), (x, a))$, where $\boldsymbol{W} := (\boldsymbol{K} + \lambda m \boldsymbol{I})^{-1}$
   Calculate the estimate $\overline{\mu}_{(x,a)} := \sum_{i=1}^{m} \alpha_i(x, a) L(x_i', \cdot)$

Algorithm 2 Estimate Value	**Algorithm 3 Approximate Value Iteration**
input Sample $\mathcal{S}$, policy π, conditional expectation estimate $\overline{\mu}_{(x,a)}$, discount γ, max. number of iterations N, error threshold θ, reward function r	**input** Sample $\mathcal{S}$, discount γ, maximum number of iterations N, reward function r, error threshold θ
output Value estimate $\hat{V}$	**output** $\overline{\mu}_{(x,a)}$, approximate optimal value $\hat{V}$
$n = 1, error = 1$; define a value vector for states $x_1', ..., x_m'$: $V := \boldsymbol{0}$	$n = 1, error = 1$; define a value vector for states $x_1', ..., x_m'$: $V := \boldsymbol{0}$
while $n \le N$ **and** $error > \theta$ **do**	Run Alg. 1 to get $\overline{\mu}_{(x,a)}$
for all $i \le m$ **do**	**while** $n \le N$ **and** $error > \theta$ **do**
$V'(x_i') \leftarrow r(x_i', \pi(x_i')) + \gamma \langle \overline{\mu}_{(x_i', \pi(x_i'))}, V \rangle_L$	**for all** $i \le m$ **do**
end for	$V'(x_i') \leftarrow \max_{a \in \mathcal{A}} r(x_i', a) + \gamma \langle \overline{\mu}_{(x_i', a)}, V \rangle_L$
$n \leftarrow n + 1$, $error \leftarrow \|V' - V\|_\infty$, $V \leftarrow V'$	**end for**
end while	$n \leftarrow n + 1$, $error \leftarrow \|V' - V\|_\infty$, $V \leftarrow V'$
return $\hat{V}(x) = r(x, \pi(x)) + \gamma \langle \overline{\mu}_{(x,\pi(x))}, V \rangle_L$	**end while**
	return $\hat{V}(x) = \max_{a \in \mathcal{A}} r(x, a) + \gamma \langle \overline{\mu}_{(x,a)}, V \rangle_L$

---

a certain contraction mapping. We now demonstrate the consistency of the estimators defined by (4) for finite state spaces, by showing that in the limit of large data the normalization of $\widehat{\alpha}$ has no effect. The following lemma is proved in the Appendix (Grünewälder et al., 2012).

Lemma 2.2. *Under the conditions of Lemma 2.1, and if $|\mathcal{X}| < \infty$ and L is strictly positive definite, by choosing $\lambda \to 0$, $\lambda^3 m \to \infty$ we have that, for any $\epsilon > 0$,*

$$\mathbb{P}_{\mathcal{S} \sim \widetilde{P}^m} \left(\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \|\mu_{(x,a)} - \widehat{\mu}_{(x,a)}\|_L > \epsilon \right) \to 0.$$

3. Application to MDPs

The learnt embeddings are applied to MDPs by recalling (4) and defining an operator

$$\widehat{\mathscr{E}}_{(x,a)}[f] := \sum_{i=1}^{m} \widehat{\alpha}_i(x, a) f(x_i'). \quad (5)$$

When $f \in \mathcal{H}_L$ we have that $\widehat{\mathscr{E}}_{(x,a)}[f] = \langle \widehat{\mu}_{(x,a)}, f \rangle_L \approx \mathbb{E}_{X \sim P(\cdot|(x,a))}[f(x)]$. When $f \notin \mathcal{H}_L$ the quality of the approximation will further depend upon how well f can be approximated by a low norm function in $\mathcal{H}_L$. This operator can be used in place of the true unknown expectation operator (2) in any MDP method which makes use of such expectations, such as dynamic programming. As an example below, we analyse value iteration, but similar considerations yield similar analyses for other methods. We summarize a joint value estimation algorithm and policy optimisation approach in the Algorithm boxes above.

If we knew P, and could efficiently compute expectations, we could define the Bellman operator B as

$$(BV)(x) := \max_{a \in \mathcal{A}} \{r(x, a) + \gamma \mathbb{E}_{X \sim P(\cdot|x,a)}[V(X)]\}, \quad (6)$$

where we suppose that the image of B is always a measurable function.[3] Recall that picking an arbitrary V_0 and iterating $V_{k+1} = BV_k$ converges in sup-norm, $V_k \to V^*$ (see e.g. Szepesvari, 2009). Since we do not know P, we use the embeddings $\widehat{\mu}_{(x,a)}$ and, recalling (5), define the operator $\widehat{B} : \mathcal{B}(\mathcal{X}) \to \mathcal{B}(\mathcal{X})$ as

$$(\widehat{B}V)(x) := \max_{a \in \mathcal{A}} \{r(x, a) + \gamma \widehat{\mathscr{E}}_{(x,a)}[V]\}. \quad (7)$$

It is necessary to define $\widehat{B}$ on functions which are not in $\mathcal{H}_L$, and this possibility introduces a term in the analysis which captures how well V^* can be approximated in $\mathcal{H}_L$ (See Thm. 3.2). By Lemma 2.2, in the limit of large data, the operator defined by (7) converges to an expectation operator on functions in $\mathcal{H}_L$, and thus $\widehat{B}$ can be seen to approximate B defined by (6) on $\mathcal{H}_L$. The following result is proved in the Appendix (Grünewälder et al., 2012):

Proposition 3.1. $\widehat{B}$ *is a sup-norm contraction on the space $\mathcal{B}(\mathcal{X})$ with Lipschitz constant γ.*

Since $\widehat{B}$ defines a sup-norm contraction mapping on a complete metric space, by Banach's fixed point theorem (e.g. Granas & Dugundji, 2003) there exists a

---

[3]We suppose for simplicity that any necessary conditions to ensure this are met, since strictly speaking B is defined only on measurable functions, see for example (Bertsekas & Shreve, 1978) for a discussion of the issues. In particular, these conditions are met when $|\mathcal{X}| < \infty$.

unique fixed point $\widehat{V}^*$ of $\widehat{B}$, such that choosing $\widehat{V}_0$ arbitrarily and iterating $\widehat{V}_{k+1} = \widehat{B}\widehat{V}_k$ converges, $\widehat{V}_k \to \widehat{V}^*$, in sup-norm,

$$||\widehat{V}_k - \widehat{V}^*||_\infty \leq \frac{\gamma^k}{1-\gamma}||\widehat{V}_1 - \widehat{V}_0||_\infty. \tag{8}$$

Suppose we perform κ iterations, obtaining the estimate $\widehat{V}_\kappa \approx \widehat{V}^*$. Once $\widehat{V}_\kappa$ is obtained we form a policy $\widehat{\pi}_\kappa$ on-the-fly[4] by acting greedily w.r.t. $\widehat{Q}_\kappa(x,a)$, where

$$\widehat{Q}_\kappa(x,a) := r(x,a) + \gamma\widehat{\mathcal{E}}_{(x,a)}[\widehat{V}_\kappa], \tag{9}$$

so that the learned policy is

$$\widehat{\pi}_\kappa(x) := \underset{a \in \mathcal{A}}{\operatorname{argmax}}\, \widehat{Q}_\kappa(x,a). \tag{10}$$

for each x in a trajectory.

Consistency: We now discuss the consistency of $\widehat{\pi}_\kappa$ as an estimate of an optimal policy π^*. The following theorem decomposes the convergence of $V^{\widehat{\pi}_\kappa}$ to the optimal value function V^*, in terms of the convergence of value iteration, the convergence of the embeddings and how well we can approximate V^* in sup-norm by a (low $||\cdot||_L$-norm) function in $\mathcal{H}_L$. This is a generic bound into which we can plug any suitable guarntees for an embedding method. In Corollary 3.3, we specialise the result to the finite state space case, where we can approximate V^* arbitrarily well.

Theorem 3.2.

$$||V^{\widehat{\pi}_\kappa} - V^*||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\Big(\gamma^\kappa||\widehat{V}_1 - \widehat{V}_0||_\infty$$
$$+ 2||V^* - \widetilde{V}^*||_\infty + \sup_{(x,a)}||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L||\widetilde{V}^*||_L\Big), \tag{11}$$

where $\widetilde{V}^*$ is any element of $\mathcal{H}_L$. Thus, whenever $\sup_{(x,a)}||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L \to 0$ in $\widetilde{P}$-probability, we have that, for any chosen $\widetilde{V}^* \in \mathcal{H}_L$,

$$||V^{\widehat{\pi}_\kappa} - V^*||_\infty \leq \frac{4\gamma}{(1-\gamma)^2}||V^* - \widetilde{V}^*||_\infty + \epsilon_\kappa + \epsilon_m, \tag{12}$$

where $\epsilon_\kappa \to 0$ and $\epsilon_m \to 0$ with convergence in $\widetilde{P}$-probability.

Proof. (Sketch, see Appendix for full proof (Grünewälder et al., 2012).) The proof hinges upon obtaining the following chain of convergences,

$$\widehat{\mathcal{E}}_{(x,a)}[\widehat{V}_\kappa] \to_{(a)} \widehat{\mathcal{E}}_{(x,a)}[\widehat{V}^*] \approx_{(b)} \widehat{\mathcal{E}}_{(x,a)}[V^*]$$
$$\approx \widehat{\mathcal{E}}_{(x,a)}[\widetilde{V}^*] \to_{(c)} \mathbb{E}_{X \sim P(\cdot|(x,a))}[\widetilde{V}^*(X)]$$
$$\approx \mathbb{E}_{X \sim P(\cdot|(x,a))}[V^*(X)].$$

---

[4]Meaning that we only need to calculate $\widehat{\pi}_\kappa(x)$ at points x in a trajectory as and when required.

The convergence (a) is a standard result for contraction mappings, (b) requires a new lemma relating the fixed points of similar contraction mappings, and (c) is possible using Lemma 2.2 because $\widetilde{V}^* \in \mathcal{H}_L$. Once this is obtained we recall that $\widehat{\pi}_\kappa$ is greedy w.r.t. $\widehat{Q}_\kappa$ defined by (9), and apply Lemma 1.1, since the optimal policy is greedy w.r.t. Q^*. □

We now interpret Thm. 3.2. The upper bound is,

$$||V^{\widehat{\pi}_\kappa} - V^*||_\infty \leq \frac{2\gamma}{(1-\gamma)^2}\Big(\overbrace{\gamma^\kappa||\widehat{V}_1 - \widehat{V}_0||_\infty}^{(i)}$$
$$+ \underbrace{2||V^* - \widetilde{V}^*||_\infty}_{(ii)} + \underbrace{\sup_{(x,a)}||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L||\widetilde{V}^*||_L}_{(iii)}\Big).$$

Here (i) is the standard difference between the value estimate of the initial policy and the value estimate of the policy that we get after applying one dynamic programming update. This term decreases to 0 with growing κ because $\gamma < 1$. (ii) is the distance from the optimal value V^* to any approximation $\widetilde{V}^*$ in the RKHS, and is therefore small when $\widetilde{V}^*$ is close to V^* and so can be smaller when $\mathcal{H}_L$ is chosen to be a richer class. Finally, (iii) measures the quality of the learned embedding: $||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L$ is the distance between the empirical estimate $\widehat{\mu}$ of the conditional distribution embedding of x' given (x,a), and the population conditional embedding μ, measured in the RKHS with kernel L. This difference is weighted by $||\widetilde{V}^*||_L$, the RKHS norm of the approximation $\widetilde{V}^*$. Intuitively, a lower RKHS norm implies a smoother function: when the norm is smaller, $\widetilde{V}^*$ is smoother, and the convergence faster. Thus (iii) requires us to obtain a better conditional mean embedding (via more training samples) when the value function is non-smooth. In other words, our approach favors smooth value functions, although given sufficient evidence, non-smooth functions can also be learned. One specialization is to the case when $V^* \in \mathcal{C}_b(\mathcal{X})$ and L is a $\mathcal{C}_b(\mathcal{X})$-universal kernel (Steinwart & Christmann, 2008, Section 4.6). In this case we can choose $\widetilde{V}^*$ such that $||V^* - \widetilde{V}^*||_\infty$ is arbitrarily small in (11).

We now specialise Thm. 3.2 to the case where $|\mathcal{X}| < \infty$ and where L is strictly positive definite kernel on $\mathcal{X}$ (we then know from Lemma 2.2 that $\sup_{(x,a)}||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L \to 0$ and that all real-valued functions are in the associated RKHS). Thus consistency is attained in otherwise very general conditions – the following is proved in the appendix:

Corollary 3.3. *Let $|\mathcal{X}| < \infty$ and L be strictly positive definite. Under the conditions of Lemma 2.2 we*
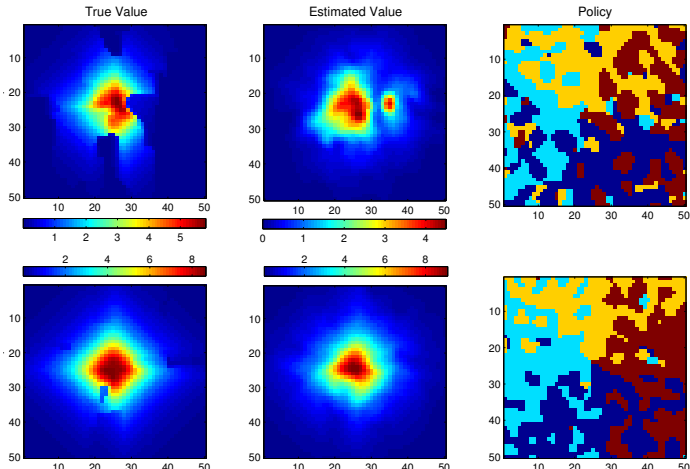
Figure 1. The left column shows the (true) value of the learned policy (color coded). The middle column shows the estimated value and the right column shows the policy. Actions are shown in the plot via a color code: yellow: go down; brown: left; dark blue: up; light blue: right. The 5000 sample policy is better (see for example the scale on the value color bars) and estimated value is close to true value. The patchy coloring is not a problem as, for example, in the bottom right it does not matter if the agent first goes up or to the left. The method has essentially learnt the task.

have that $||V^{\widehat{\pi}_\kappa} - V^*||_\infty \to 0$ *with convergence in* $\widetilde{P}$-*probability.*

Complexity analysis: Once the embeddings are learnt, the complexity of learning the approximate value function $\widehat{Q}_\kappa$ is $\mathcal{O}(m^2|\mathcal{A}|\kappa)$: due to the expansion of $\widehat{\mu}_{(x,a)}$ in the m points in $\mathcal{S}$, computing each expectation is $\mathcal{O}(m)$ and we only ever need to know the evaluation of each iterate $\widehat{V}_k$ at the m points in $\mathcal{S}$. Applying the learnt policy (10) to a trajectory $(x_0, x_1, \ldots, x_T)$ of length T, is similarly $\mathcal{O}(m|\mathcal{A}|T)$. In Sec. B of the Supplementary material, we propose a sparser representation of the embedding, using an incomplete Cholesky approximation (Shawe-Taylor & Cristianini, 2004)[Sec. 5.2]. This reduces the cost of learning the embeddings from cubic to linear in m, and allows us to compute subsequent expectations in $\mathcal{O}(\ell)$, where generally $\ell \ll m$.

4. Experiments

We performed three experiments, using the embeddings in value estimation and policy optimization. The first experiment was an MDP with a fully observed discrete state space, to demonstrate convergence of the value function with increasing training sample size. The second and third experiments evaluate our approach on a classical control task and a task with high dimensional states. In policy optimisation we compare to LSPI (Lagoudakis & Parr, 2003) where we use the q-value estimator from (Engel et al., 2005), and for value estimation we compare to NPDP[5] (Kroemer & Peters, 2011). We achieve better performance in all our experiments.

We briefly address the choice of the regularization term λ. It can be shown that the conditional embeddings

---

[5] We thank the authors for providing code.

solve,

$$\hat{\mu} := \underset{\mu \in \mathcal{H}}{\operatorname{argmin}} \left[\sum_{i=1}^{m} \|L(x_i', \cdot) - \mu(x_i, a_i)\|_L^2 + \lambda \|\mu\|_{\mathcal{H}}^2 \right].$$

where $\mathcal{H} \subseteq (\mathcal{H}_L)^{(\mathcal{X} \times \mathcal{A})}$, recovering the vector-valued regression setting of Micchelli & Pontil (2005) (see Sec. D for details) which provides cross validation scheme for the parameter λ.

4.1. Experiment 1

The first experiment is a navigation experiment in a 50 x 50 room. The reward is a Gaussian centered in the middle of the room. The agent has four actions: go north, east, south or west. Each action has a success rate of 80 % and results in random movement with 20 % chance. The state space is fully observed. We learn the conditional distribution embedding from either 1000 or 5000 uniformly sampled transitions, uniformity ensuring we avoid exploration artifacts. We used a Gaussian kernel and cross-validated to determine the regulariser. Results are shown in Figure 1.

4.2. Experiment 2

We consider the under-actuated pendulum swing up task (Deisenroth et al., 2009). We generate a discrete-time approximation of the continuous-time pendulum dynamics as done in (Deisenroth et al., 2009). Starting from an arbitrary state the goal is to swing the pendulum up and balance it in the inverted position. The applied torque is $u \in [-5, 5]Nm$ and is not sufficient for a direct swing up. The state space is defined by the angle $\theta \in [-\pi, \pi]$ and the angular velocity, $\omega \in [-7, 7]$. The reward is given by the function $r(\theta, \omega) = \exp(-\theta^2 - 0.2\omega^2)$. For policy learning we compared to the GP-based LSPI approach and for value learning to NPDP. The results of the comparison are shown in Fig. 2.
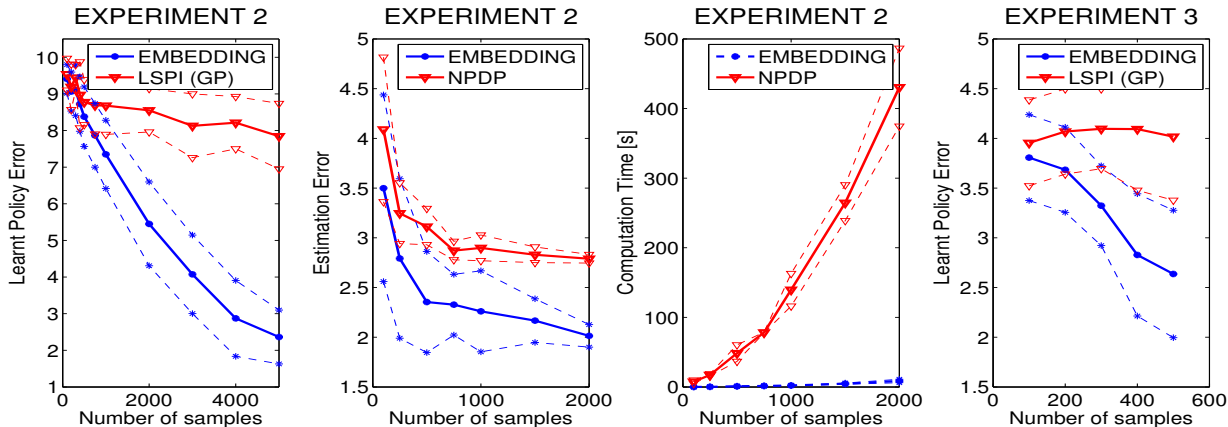
Figure 2. We compared our approach in policy learning to the GP based LSPI approach (1st and 4th plot) and in value learning to NPDP (2nd and 3rd plot). 1st and 4th plot: y-axis shows the average difference of the value of the learnt policy to the optimal value (averaged over the state space). In experiment 2 LSPI improves slowly after 500 samples and has problems with the two room task for small samples. 2nd plot: y-axis shows the prediction error of the value estimators (averaged over the state space). The embedding-based estimator is significantly better, especially for higher sample numbers. 3rd plot: run time for the two methods – the embedding method is 50-110 times faster on this task.

Details for the policy learning setting: We sampled uniformly from the state and action space and used a Gaussian kernel on both, selecting as kernel width the average K-neighbour distance, where K is one quarter of the sample size. We considered a discretization of the action space into 25 actions and we measured the difference between the value function evaluated on a grid of 25×25 points to the optimal value obtained by dynamic programming using the deterministic system dynamics. We compared over different sample sizes and averaged the performance over 10 repetitions.

Details for the value estimation setting: We used the optimal policy to generate samples. The goal was to predict the value of the optimal policy. The performance of NPDP depends strongly on the bandwidth parameter of the used kernel (a Gaussian). For parameter selection, we optimised performance on a validation set over a grid all free parameters (bandwidth for NPDP, bandwidth and λ for the embedding), and report the error on an independent test set. The relatively poor scaling of NPDP with increased sample size is due to the numerical integration step in (Kroemer & Peters, 2011, Algorithm 1).

4.3. Experiment 3

Our final experiment is a high dimensional task where sensor measurements are available, and no state description is present. The environment consists of two rooms connected via a short corridor (Böhmer, 2012). The sensor measurements are images from a 3D renderer, and we aggregate four orientations (north, east, south and west) for a panorama, since the camera im-
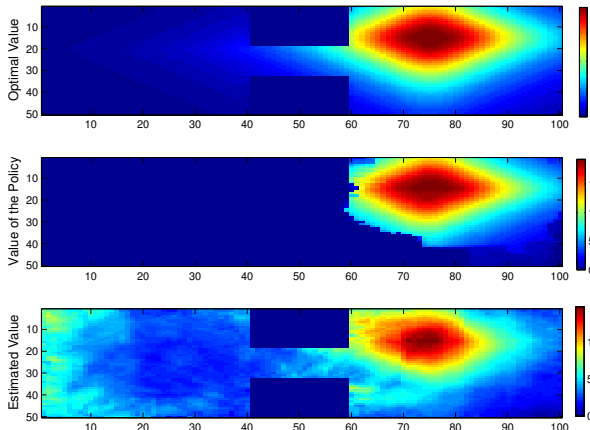

Figure 3. Results for Experiment 3 with 4000 data points. Top figure: optimal value. Middle figure: value of the learned policy. Bottom figure: the predicted value. The policy is nearly perfect in the room containing the objective. The performance degrades when the corridor is reached due to the challenging ambiguous nature of the images, which are insufficient to accurately distinguish between the locations. Similarly, the left wall has a high predicted value. The bottom picture shows that the value estimate is close to the optimal value.

ages are ambiguous, especially close to the walls. The task of the agent is to reach a goal located in one of the rooms, using only the images to orient itself. Training points were chosen uniformly over the input space. We used a Gaussian kernel and cross-validated the regularization parameter. Results for 4000 training points are shown in Figure 3. We compared to the GP based LSPI approach using the same kernel and settings for both approaches; results are shown in Figure 2. Our method improves with increasing sample

numbers. The GP based LSPI approach has obvious difficulties with this task and does not improve. We did not apply NPDP to exp. 3, as it would be computationally intractable in given the high dimensionality.

5. Conclusions and Outlook

We have proposed a novel application of RKHS embeddings to learning expectation operators associated to transition dynamics in MDPs, with particular focus on their use in dynamic programming methods. The approach avoids the need for density estimates, sampling methods for evaluation of integrals, or explicit models of the system; is computationally efficient, having cost linear in the number of samples used in training (or even sublinear, with appropriate approximations); and has performance guarantees. Future work will focus on generalizing to more complex state and action spaces, and extending the convergence results to continuous state spaces. Another important generalization concerns the sampling distribution, which here is assumed to be iid, but one can expect similar results to hold in the non-iid case.

Acknowledgements

References

Bach, F. R. and Jordan, M. I. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2002.

Bertsekas, D. and Shreve, S. Mathematical issues in dynamic programming. 1978. Supplementary material to Stochastic Optimal Control: The Discrete-Time Case, Academic press, 1978.

Böhmer, W. Robot Navigation using Reinforcement Learning and Slow Feature Analysis. *ArXiv*, 2012.

Deisenroth, M. P., Rasmussen, C. E., and Peters, J. Gaussian process dynamic programming. *Neurocomputing*, 72 (7-9):1508–1524, 2009.

Engel, Y., Mannor, S., and Meir, R. Reinforcement learning with gaussian processes. In *ICML*, 2005.

Granas, A. and Dugundji, J. *Fixed Point Theory*. Springer-Verlag, 2003.

Grünewälder, S., Lever, G., Baldassarre, L., Pontil, M., and Gretton, A. Modelling transition dynamics in mdps with rkhs embeddings. In *arXiv*, 2012.

Kroemer, O. and Peters, J. A non-parametric approach to dynamic programming. In *NIPS*, 2011.

Lagoudakis, M. G. and Parr, R. Least-squares policy iteration. *J. Mach. Learn. Res.*, 4, December 2003.

Micchelli, C.A. and Pontil, M. On learning vector-valued functions. *Neural Computation*, 17(1):177–204, 2005.

Ormoneit, D. and Sen, S. Kernel-based reinforcement learning. In *Machine Learning*, pp. 161–178, 1999.

Parr, R., Li, L., Taylor, G., Painter-Wakefield, C., and Littman, M. L. An analysis of linear models, linear value-function approximation, and feature selection for reinforcement learning. In *ICML*, 2008.

Petersen, K. B. and Pedersen, M. S. The matrix cookbook, 2008.

Rasmussen, C. E. and Kuss, M. Gaussian processes in reinforcement learning. In *NIPS*. MIT Press, 2003.

Reed, M. and Simon, B. *Methods of modern mathematical physics. Vol. 1: Functional Analysis*. Academic Press, San Diego, 1980.

Shawe-Taylor, J. and Cristianini, N. *Kernel Methods for Pattern Analysis*. Cambridge University Press, 2004.

Singh, S. P. and Yee, R. C. An upper bound on the loss from approximate optimal-value functions. *Machine Learning*, 16(3):227–233, 1994.

Smola, A. J. and Kondor, R. I. Kernels and regularization on graphs. In *COLT*, 2003.

Song, L., Huang, J., Smola, A. J., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *ICML*, 2009.

Song, L., Boots, B., Siddiqi, S., Gordon, G., and Smola, A. Hilbert space embeddings of hidden Markov models. In *ICML*, 2010a.

Song, L., Gretton, A., and Guestrin, C. Nonparametric tree graphical models. *AISTATS*, 9, 2010b.

Song, L., Gretton, A., Bickson, D., Low, Y., and Guestrin, C. Kernel belief propagation. In *AISTATS*, 2011.

Sriperumbudur, B., Gretton, A., Fukumizu, K., Lanckriet, G., and Schölkopf, B. Hilbert space embeddings and metrics on probability measures. *JMLR*, 11, 2010.

Steinwart, I. and Christmann, A. *Support Vector Machines*. Springer, 2008.

Szepesvari, C. *Algorithms for reinforcement learning*. Morgan and Claypool, 2009.

Taylor, G. and Parr, R. Kernelized value function approximation for reinforcement learning. In *ICML*, 2009.

Wasserman, L. *All of Nonparametric Statistics*. Springer, 2006.

Wendland, H. *Scattered Data Approximation*. Cambridge University Press, Cambridge, UK, 2005.

Xu, X., Xie, T., Hu, D., and Lu, X. Kernel least-squares temporal difference learning. *International Journal of Information Technology*, 2005.

A. Approximate Policy Iteration

Algorithm 4 Approximate Policy Iteration

---

input Sample $\mathcal{S}$, discount γ, maximum number of iterations N, reward function r
output $\overline{\mu}_{(x,a)}$, optimised policy π and value V on $x'_1, ..., x'_m$
  $n = 1$, $\pi' = \mathbf{0}$
  Initialise a random policy π,
  Run Alg. 1 to get $\overline{\mu}_{(x,a)}$
  while $n \leq N$ **and** $\pi' \neq \pi$ **do**
    $\pi' \leftarrow \pi$, $n \leftarrow n+1$, Run Alg. 2 to get V for π
    for all $i \leq m$ **do**
      $\pi(x'_i) \leftarrow \arg\max_{a \in \mathcal{A}} r(x'_i, a) + \gamma \langle \overline{\mu}_{(x'_i, a)}, V \rangle_L$
    end for
  end while

B. Complexity improvements via sparse representations

We propose a low rank approximation to efficiently compute the conditional expectation of the reward, assuming this is in an RKHS, and following the approach of (Song et al., 2011)[Sec. 5]. For ease of notation, we define the current state x, the next state x', and the action a. If we assume V is an RKHS function, our estimate of the expected conditional value can be written as

$$\mathbb{E}_{\overline{\mu}_{(x,a)}}[V] = \sum_{i=1}^{m} \alpha_i(x, a) V(x'_i) = \sum_{i=1}^{m} \alpha_i(x, a) \langle V(x'), l(x, x'_i) \rangle = \langle V, \Phi\alpha \rangle$$

where α is a vector with ith entry $\alpha_i(x, a)$, and $\Phi := \begin{bmatrix} l(x'_1, \cdot) & \ldots & l(x'_m, \cdot) \end{bmatrix}$. We expand $V = \Phi\beta$, since any component of V orthogonal to the span of Φ will project to zero. We further define $\Upsilon := \begin{bmatrix} k((x_1, a_1), \cdot) & \ldots & k((x_m, a_m), \cdot) \end{bmatrix}$, the embedding of the training sample of state-action pairs used in obtaining α. We define the Gram matrices $L := \Phi^\top \Phi$ and $K := \Upsilon^\top \Upsilon$ as before. Following (Shawe-Taylor & Cristianini, 2004)[Sec. 5.2], we may approximate $\Phi \approx QR$ using an incomplete Gram-Schmidt procedure, where Q contains ℓ orthonormal columns spanning a subspace of Φ, and $R \in \mathbb{R}^{\ell \times m}$ is upper triangular in its first ℓ columns, assuming the columns of Φ are in the order of elimination in the Gram Schmidt procedure. Generally, $\ell \ll m$ yields a good approximation to Φ: see (Bach & Jordan, 2002)[Appendix C]. In the same way, we can define $\Upsilon \approx GH$, where G is a set of orthonormal functions spanning a subspace of Υ, H has dimension $\ell_2 \times m$, and $\ell_2 \ll m$. Denote by $\mathcal{I}$ the index set of the first ℓ columns of Φ, and by $\mathcal{J}$ the indices of the first ℓ_2 columns of Υ. Then

$$\Phi_{\mathcal{I}} = QR_{\mathcal{I}} \quad \text{hence} \quad Q = \Phi_{\mathcal{I}} R_{\mathcal{I}}^{-1},$$

from which $\Phi \approx QR = \Phi_{\mathcal{I}} R_{\mathcal{I}}^{-1} R =: \Phi_{\mathcal{I}} A$, where $A \in \mathbb{R}^{\ell \times m}$. Similarly, $\Upsilon \approx \Upsilon_{\mathcal{J}} (H_{\mathcal{J}})^{-1} H =: \Upsilon_{\mathcal{J}} B$. Given that

$$\alpha(\cdot) = (K + \lambda m I)^{-1} \Upsilon^\top k((x, a), \cdot),$$

the expected value is

$$\begin{aligned}
\mathbb{E}_{\overline{\mu}_{(x,a)}}[V] &= \left\langle \Phi(K + \lambda m I)^{-1} \Upsilon^\top k((x, a), \cdot), \Phi\beta \right\rangle \\
&\approx \left\langle \Phi_{\mathcal{I}} A \left(B^\top \Upsilon_{\mathcal{J}}^\top \Upsilon_{\mathcal{J}} B + \lambda m I \right)^{-1} B^\top \Upsilon_{\mathcal{J}}^\top k((x, a), \cdot), \Phi_{\mathcal{I}} A\beta \right\rangle \\
&= \beta^\top A^\top L_{\mathcal{I}} A B^\top \left(BB^\top + \lambda m K_{\mathcal{J}}^{-1} \right)^{-1} K_{\mathcal{J}}^{-1} k_{\mathcal{J}}(x, a)
\end{aligned}$$

where $L_{\mathcal{I}}$ and $K_{\mathcal{J}}$ are the submatrices indexed by $\mathcal{I}$ and $\mathcal{J}$, respectively, $k_{\mathcal{J}}(x, a)$ is a vector of kernels between the set $\mathcal{J}$ of retained points and the new observation (x, a), and we use the form of the Woodbury identity for positive definite matrices in (Petersen & Pedersen, 2008)[eq. (147) p. 17] for the final line. A nice feature of this approach is that we only need evaluate the inner product of a new state-action pair with the points in the index set $\mathcal{J}$, which allows us to rapidly compute the expected value function over a wide range of state-action pairs at reasonable cost.

C. Proofs

C.1. Auxiliary results

Theorem C.1. *(e.g. Smola & Kondor, 2003, Thm. 4) Given a finite set of points $\mathcal{Z} = \{z_1, \ldots, z_t\}$, consider $\boldsymbol{h} \in \mathbb{R}^{\mathcal{Z}}$ as a vector $\boldsymbol{h} \in \mathbb{R}^t$ via $\boldsymbol{h}(z_i) := \boldsymbol{h}^\top \boldsymbol{e}_i = h_i$. Then, for a symmetric, p.s.d. matrix $\boldsymbol{R}$ the Hilbert space $\mathcal{H} = \mathrm{im}(\boldsymbol{R}) \subseteq \mathbb{R}^t$ of real-valued functions on $\mathcal{Z}$ with inner product $\langle \boldsymbol{h}, \boldsymbol{g} \rangle_{\mathcal{H}} = \boldsymbol{h}^\top \boldsymbol{R} \boldsymbol{g}$ is an RHKS whose reproducing kernel $K : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ is given by the pseudoinverse $\boldsymbol{R}^+$, i.e. such that $K(z_i, z_j) := R_{ij}^+ = \boldsymbol{e}_i^\top \boldsymbol{R}^+ \boldsymbol{e}_j$.*

C.2. Proof of Lemma 1.1

The proof is due to (Singh & Yee, 1994).

Proof. For any action-value function $Q : \mathcal{X} \times \mathcal{A} \to \mathbb{R}$ and any x let $a^* = \mathrm{argmax}_{a \in \mathcal{A}} Q^*(x, a)$, $\widehat{a} = \mathrm{argmax}_{a \in \mathcal{A}} Q(x, a)$. Suppose that $||Q^* - Q||_\infty \le \epsilon$. We have $Q^*(x, a^*) - \epsilon \le Q(x, a^*) \le Q(x, \widehat{a}) \le Q^*(x, \widehat{a}) + \epsilon$. Therefore,

$$r(x, a^*) + \gamma \mathbb{E}_{X \sim P(\cdot | x, a^*)}[V^*(X)] - \epsilon \le r(x, \widehat{a}) + \gamma \mathbb{E}_{X \sim P(\cdot | x, \widehat{a})}[V^*(X)] + \epsilon$$
$$r(x, a^*) + \gamma \mathbb{E}_{X \sim P(\cdot | x, a^*)}[V^*(X)] - r(x, \widehat{a}) \le \gamma \mathbb{E}_{X \sim P(\cdot | x, \widehat{a})}[V^*(X)] + 2\epsilon.$$

Now,

$$V^*(x) - V^{\pi_Q}(x) = r(x, a^*) + \gamma \mathbb{E}_{X \sim P(\cdot | x, a^*)}[V^*(X)] - r(x, \widehat{a}) - \gamma \mathbb{E}_{X \sim P(\cdot | x, \widehat{a})}[V^{\pi_Q}(X)]$$
$$\le \gamma \mathbb{E}_{X \sim P(\cdot | x, \widehat{a})}[V^*(X)] - \gamma \mathbb{E}_{X \sim P(\cdot | x, \widehat{a})}[V^{\pi_Q}(X)] + 2\epsilon$$
$$\le \gamma \sup_{x' \in \mathcal{X}} |V^*(x') - V^{\pi_Q}(x')| + 2\epsilon$$
$$\sup_{x \in \mathcal{X}} |V^*(x) - V^{\pi_Q}(x)| \le \frac{2\epsilon}{1 - \gamma}.$$

$\square$

C.3. Proof of Lemma 2.1

Proof. Following (Song et al., 2009) there exists a *conditional embedding operator* $U : \mathcal{H}_K \to \mathcal{H}_L$ such that $U(K((x, a), \cdot)) = \mu_{(x,a)}$ for all $(x, a) \in \mathcal{X} \times \mathcal{A}$. From (Song et al., 2010b) there exists an empirical embedding operator $\overline{U} : \mathcal{H}_K \to \mathcal{H}_L$ such that $\overline{\mu}_{(x,a)} = \overline{U}(K((x, a), \cdot))$ and,

$$\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} \{||\mu_{(x,a)} - \overline{\mu}_{(x,a)}||_L\} = \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} ||(U - \overline{U})(K((x, a), \cdot))||_L$$
$$\le \sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} ||K((x, a), \cdot)||_K ||(U - \overline{U})||_{\mathrm{HS}}$$
$$= \rho_K ||(U - \overline{U})||_{\mathrm{HS}},$$

where $||\cdot||_{\mathrm{HS}}$ denotes the Hilbert-Schmidt norm, and the result follows directly from (Song et al., 2010b)[Thm. 1].

$\square$

C.4. Proof of Lemma 2.2

Proof. Since L is strictly positive definite on a finite set (Smola & Kondor, 2003, Thm. 4) (reproduced in the appendix as Thm. C.1) demonstrates that $\mathcal{H}_L$ consists of all bounded functions on $\mathcal{X} = \{x_1, \ldots, x_n\}$ and $\langle f, g \rangle_L = \boldsymbol{f}^\top \boldsymbol{L}^{-1} \boldsymbol{g}$ where $\boldsymbol{L} = (L(x_i, x_j))_{i,j=1}^n$ and $\boldsymbol{f} := (f(x_i))_{i=1}^n$ is the vector of point evaluations of f on $\mathcal{X}$. Thus $\mathcal{H}_L$ contains all binary functions on $\mathcal{X}$, $\mathrm{bin}(\mathcal{X}) := \{-1, 1\}^{\mathcal{X}} \subset \mathcal{H}_L$, and let $b := \max_{f \in \mathrm{bin}(\mathcal{X})} ||f||_L < \infty$.

Now, given any $\epsilon > 0$, Lemma 2.1 implies that, uniformly over all (x, a),

$$
\mathbb{P}_{\mathcal{S} \sim \widetilde{P}} \left(\max_{f \in \text{bin}(\mathcal{X})} |\mathbb{E}_{X \sim P(\cdot|(x,a))}[f(X)] - \sum_{i=1}^m \alpha_i(x, a) f(x_i')| > \epsilon \right)
$$

$$
= \mathbb{P}_{\mathcal{S} \sim \widetilde{P}} \left(\max_{f \in \text{bin}(\mathcal{X})} |\langle \mu_{(x,a)} - \overline{\mu}_{(x,a)}, f \rangle_L| > \epsilon \right)
$$

$$
\leq \mathbb{P}_{\mathcal{S} \sim \widetilde{P}} \left(||\mu_{(x,a)} - \overline{\mu}_{(x,a)}||_L b > \epsilon \right)
$$

$$
\to 0. \tag{13}
$$

Consideration of the all ones function, $f : x \mapsto 1 \in \text{bin}(\mathcal{X})$, in (13) implies that

$$
\mathbb{P}_{\mathcal{S} \sim \widetilde{P}} \left(\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |1 - \sum_{i=1}^m \alpha_i(x, a)| > \epsilon \right) \to 0. \tag{14}
$$

Consideration of the function $g(x_i) = -1$ if $\alpha_i(x, a) < 0$ and which is equal to $+1$ otherwise, in (13), implies that

$$
\mathbb{P}_{\mathcal{S} \sim \widetilde{P}} \left(\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |c - \sum_{i=1}^m |\alpha_i(x, a)|| > \epsilon \right) \to 0, \tag{15}
$$

where $c = \mathbb{E}_{X \sim P(\cdot|(x,a))}[g(X)] \leq 1$. Together (14) and (15) imply that

$$
\mathbb{P}_{\mathcal{S} \sim \widetilde{P}} \left(\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} |1 - \sum_{i=1}^m |\alpha_i(x, a)|| > \epsilon \right) \to 0.
$$

This implies that $\sup_{\{i,(x,a)\}} |\widehat{\alpha}_i(x, a) - \alpha_i(x, a)| \to 0$ in $\widetilde{P}$-probability. And since $||\widehat{\mu}_{(x,a)} - \overline{\mu}_{(x,a)}||_L^2 = \sum_{ij}(\widehat{\alpha}_i(x, a) - \alpha_i(x, a)) L_{ij}^{-1} (\widehat{\alpha}_j(x, a) - \alpha_j(x, a))$ this implies that, $\sup_{(x,a) \in \mathcal{X} \times \mathcal{A}} ||\widehat{\mu}_{(x,a)} - \overline{\mu}_{(x,a)}||_L \to 0$ in $\widetilde{P}$ probability and the result follows from the consistency of $\overline{\mu}_{(x,a)}$, uniformly over (x, a), Lemma 2.1. $\qquad \square$

C.5. Proof of Proposition 3.1

Proof. for any $V, W \in \mathcal{B}(\mathcal{X})$ we have,

$$
\sup_{x \in \mathcal{X}} |(\widehat{B}V)(x) - (\widehat{B}W)(x)| = \sup_{x \in \mathcal{X}} |\max_{a \in \mathcal{A}}\{r(x, a) + \gamma \widehat{\mathscr{E}}_{(x,a)}[V]\} - \max_{a \in \mathcal{A}}\{r(x, a) + \gamma \widehat{\mathscr{E}}_{(x,a)}[W]|\}
$$

$$
\leq \sup_{(x,a)} |\gamma \widehat{\mathscr{E}}_{(x,a)}[V - W]|
$$

$$
\leq \gamma \sup_{x \in \mathcal{X}} |V(x) - W(x)|.
$$

$\qquad \square$

C.6. Proof of Thm. 3.2

We will require the following lemma regarding contraction mappings:

Lemma C.2. *Suppose that $C, \widehat{C}$ are contraction mappings on Banach spaces $(\mathcal{V}, ||\cdot||)$, $(\widehat{\mathcal{V}}, ||\cdot||)$ with $\mathcal{V} \subseteq \widehat{\mathcal{V}}$ and with fixed points $v^*, \hat{v}^*$ respectively. Suppose that $||(C - \widehat{C})(v^*)|| < \epsilon$ and that $||\widehat{C}(v) - \widehat{C}(w)|| < \gamma ||v - w||$, then $||v^* - \hat{v}^*|| \leq \frac{\epsilon}{1-\gamma}$.*

Proof. There exist sequences $v_n \to v^*$, $\hat{v}_n \to \hat{v}^*$ where $v_{n+1} = C(v_n)$, $\hat{v}_{n+1} = \widehat{C}(\hat{v}_n)$ so that $||\hat{v}_{n+1} - v_{n+1}|| = ||\widehat{C}(\hat{v}_n) - C(v_n)|| \leq ||\widehat{C}(\hat{v}_n) - \widehat{C}(v_n)|| + ||(\widehat{C} - C)(v_n)|| \leq \gamma ||\hat{v}_n - v_n|| + ||(\widehat{C} - C)(v_n)||$. Taking the limit $n \to \infty$, and the continuity of $(\widehat{C} - C)$ implies $||\hat{v}^* - v^*|| \leq \gamma ||\hat{v}^* - v^*|| + ||(\widehat{C} - C)(v^*)|| \leq \gamma ||\hat{v}^* - v^*|| + \epsilon$. $\qquad \square$

Corollary C.3. $||V^* - \widehat{V}^*||_\infty \leq \gamma \frac{2||V^* - \widetilde{V}^*||_\infty + \sup_{(x,a)} ||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L ||\widetilde{V}^*||_L}{1-\gamma}$ *where $\widetilde{V}^*$ is any element of $\mathcal{H}_L$.*

Proof. We have that,

$$||(B - \widehat{B})(V^*)||_\infty := \sup_{x \in \mathcal{X}} |\max_{a \in \mathcal{A}} \{r(x,a) + \gamma \mathbb{E}_{X \sim P(\cdot|x,a)}[V^*(X)]\} - \max_{a \in \mathcal{A}} \{r(x,a) + \gamma \widehat{\mathscr{E}}_{(x,a)}[V^*]\}|$$

$$\leq \gamma \sup_{(x,a)} |\mathbb{E}_{X \sim P(\cdot|x,a)}[V^*(X)] - \widehat{\mathscr{E}}_{(x,a)}[V^*]|$$

$$\leq \gamma \sup_{(x,a)} \left(|(\mathbb{E}_{X \sim P(\cdot|x,a)} - \widehat{\mathscr{E}}_{(x,a)})[V^* - \widetilde{V}^*]| + |\langle \mu_{(x,a)} - \widehat{\mu}_{(x,a)}, \widetilde{V}^* \rangle_L| \right)$$

$$\leq \gamma \left(2||V^* - \widetilde{V}^*||_\infty + \sup_{(x,a)} ||\mu_{(x,a)} - \widehat{\mu}_{(x,a)}||_L ||\widetilde{V}^*||_L \right),$$

and the result follows from Lemma C.2. $\qquad \square$

We can now prove Thm. 3.2:

Proof. We recall definition (1) and (9), and we have,

$$||Q^* - \widehat{Q}_\kappa||_\infty = \sup_{(x,a)} |\widehat{Q}_\kappa(x,a) - Q^*(x,a)|$$

$$\leq \gamma \sup_{(x,a)} |\widehat{\mathscr{E}}_{(x,a)}[\widehat{V}_\kappa(X)] - \mathbb{E}_{X \sim P(\cdot|x,a)}[V^*(X)]|$$

$$\leq \gamma \left(\sup_{(x,a)} |\widehat{\mathscr{E}}_{(x,a)}[\widehat{V}_\kappa - \widehat{V}^*]| + \sup_{(x,a)} |\widehat{\mathscr{E}}_{(x,a)}[\widehat{V}^* - V^*]| \right.$$

$$+ \sup_{(x,a)} |(\widehat{\mathscr{E}}_{(x,a)} - \mathbb{E}_{X \sim P(\cdot|x,a)})[V^* - \widetilde{V}^*]|$$

$$\left. + \sup_{(x,a)} |(\widehat{\mathscr{E}}_{(x,a)} - \mathbb{E}_{X \sim P(\cdot|x,a)})[\widetilde{V}^*]| \right), \tag{16}$$

where $\widetilde{V}^*$ is any element of $\mathcal{H}_L$. Continuing,

$$(16) \leq \gamma \left(||\widehat{V}_\kappa - \widehat{V}^*||_\infty + ||\widehat{V}^* - V^*||_\infty + 2||V^* - \widetilde{V}^*||_\infty + \sup_{(x,a)} |\langle \mu_{(x,a)} - \widehat{\mu}_{(x,a)}, \widetilde{V}^* \rangle_L| \right)$$

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

where the final line follows by applying (8) and Corollary C.3 to the first and second terms of the preceding line. Lemma 1.1 then implies that,

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

proving (11), and (12) follows by applying Lemma 2.2. $\qquad \square$

C.7. Proof of Corollary 3.3

Proof. Recalling Thm. C.1, since L is strictly positive definite kernel on a finite set the RHKS $\mathcal{H}_L$ consists of all bounded real-valued functions on $\mathcal{X}$ so we can choose $\widetilde{V}^* = V^*$ in (12). $\qquad \square$

D. Kernel conditional embeddings as solution to regression problem

We demonstrate that kernel conditional mean embedding can be obtained as the solution to a regression problem, where the output space is an RKHS and the loss is described by the associated RKHS norm. Our discussion is in general terms, rather than using the specific spaces required in the MDP setting.

Let $\mathcal{F}$ be an RKHS on the set $\mathcal{X}$ with kernel K and feature map $\phi(x)$, and $\mathcal{G}$ be an RKHS on $\mathcal{Y}$ with kernel L and feature map $\varphi(y)$. Consider the problem of finding a mapping $f : \mathcal{X} \to \mathcal{G}$ such that,

$$\mathbb{E}_X \left(\mathbb{E}_{Y \sim P(\cdot|X)}[g] - \langle f(x), g \rangle_{\mathcal{G}} \right)^2, \tag{17}$$

is small over *all* g in some ball $\widetilde{\mathcal{G}} \subset \mathcal{G}$. This is clearly minimized (to zero) when $f(x)$ is the true embedding $\mu(x)$ of $P(\cdot|x)$ in $\mathcal{G}$. In particular,

$$\begin{aligned}
\mathbb{E}_X \left(\mathbb{E}_{Y \sim P(\cdot|X)}[g] - \langle f(x), g \rangle_{\mathcal{G}} \right)^2 &= \mathbb{E}_X \left\langle g, \mu_{(X)} - f(X) \right\rangle_{\mathcal{G}}^2 \\
&\leq \mathbb{E}_X \left\| \mu_{(X)} - f(X) \right\|_{\mathcal{G}}^2 \|g\|_{\mathcal{G}}^2 \\
&\leq \mathbb{E}_{XY} \left\| L(Y, \cdot) - f(X) \right\|_{\mathcal{G}}^2 \|g\|_{\mathcal{G}}^2
\end{aligned}$$

where we successively use Cauchy-Schwarz and Jensen's inequality. Thus, minimizing

$$E_{XY} \| L(Y, \cdot) - \mu(X) \|_{\mathcal{G}}^2 \tag{18}$$

is a good surrogate to the objective (17) over balls in $\mathcal{G}$, and should give better guarantees on "smoother" functions g (that is, low norm functions).

There are two ways to think about this problem: either μ is a mapping between $\phi(x) \in \mathcal{F}$ and $\mathcal{G}$, i.e., a mapping between feature spaces, in which case the solution is regularized least squares regression in feature space (Sec. D.1); or μ maps directly from $\mathcal{X}$ to $\mathcal{G}$, and the solution is a special case of the approach of (Micchelli & Pontil, 2005) (Sec. D.2).

D.1. Regression between feature spaces

To describe the solution to the regression problem (18), we must first introduce the tensor space of functions $\mathcal{F} \otimes \mathcal{G}$, which is a subspace of $\mathcal{H} := \mathcal{G}^{\mathcal{F}}$. The tensor product in Hilbert spaces is an immediate extension of the outer product $ab^\top$ of vectors $a \in \mathbb{R}^d$, $b \in \mathbb{R}^{d'}$, and is described e.g. in (Reed & Simon, 1980)[Ch. 2.4]. The inner product between two elements of $\mathcal{F} \otimes \mathcal{G}$ is

$$\langle f_1 \otimes g_1, f_2 \otimes g_2 \rangle_{\mathrm{HS}} := \langle f_1, f_2 \rangle_{\mathcal{F}} \langle g_1, g_2 \rangle_{\mathcal{G}}.$$

Note that the tensor inner product is defined in terms of the inner products on $\mathcal{F}$ and $\mathcal{G}$, which must therefore be Hilbert spaces (although $\mathcal{X}$ and $\mathcal{Y}$ need only be sets). The action of the operator $g \otimes f \in \mathcal{H}$ is

$$(g \otimes f_1) f_2 := g \langle f_1, f_2 \rangle.$$

We will be solving a regression problem which maps $\phi(x)$ to $\varphi(y)$. The link with conditional mean embeddings is immediate from the map from $\mathcal{F} \to \mathcal{G}$, since the conditional mean embedding is a map from feature space to feature space. For instance, if $\mathcal{X}$ is a finite set of size d (as is $\mathcal{Y}$), then it has the natural feature map $\phi(x) := e_x$, where e_s is the unit vector in $\mathbb{R}^p$ taking value 1 in dimension s, and zero in the remaining entries. In this case, the conditional mean embedding is the conditional probability, and the conditional mean is indeed a mapping from the feature space $\mathcal{F}$ to the feature space $\mathcal{G}$: see (Song et al., 2010a)[Sec. 3.3] for details. A second special case is where $\mathcal{X} := \mathcal{F}$, $\mathcal{Y} := \mathcal{G}$, and x and y are jointly Gaussian distributed with zero mean. In this case, the conditional mean embedding is the conditional mean of y given x, or in other words, the solution to a standard least squares regression problem (which may or may not be regularized). Note that in this case, we rely on $\mathcal{X}$ and $\mathcal{Y}$ being Hilbert spaces and having well defined inner products, which allows us to identify them with $\mathcal{F}$ and $\mathcal{G}$, respectively. This will not necessarily be true in general.

We next introduce the (uncentered) cross-covariance operator, C_{XY}. This belongs to the space $\mathrm{HS}(\mathcal{G}, \mathcal{F})$ of Hilbert-Schmidt operators mapping from $\mathcal{G}$ to $\mathcal{F}$, and is defined such that

$$\langle f, C_{XY} g \rangle_{\mathcal{F}} := \mathrm{cov}(f(X), g(Y)) = \mathbb{E}_{XY}[f(X)g(Y)]$$

for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$. The uncentered covariance operator $C_{XX} \in \mathrm{HS}(\mathcal{F}, \mathcal{F})$ may be defined by analogy: for all $f_1, f_2 \in \mathcal{F}$,

$$\langle f_1, C_{XX} f_2 \rangle_{\mathcal{F}} := \mathbb{E}_X[f_1(X) f_2(X)].$$

For more details, see e.g. (Song et al., 2009; 2010b) and references therein.

We now turn to the minimization of problem (18). Define the operator $A : \mathcal{F} \to \mathcal{G}$ (assumed to be Hilbert-Schmidt). Then the problem of minimizing (18) can be written,

$$\min_A \left(\mathbb{E}_{XY} \left\| \varphi(y) - A\phi(x) \right\|_{\mathcal{G}}^2 + \lambda \left\| A \right\|_{HS}^2 \right). \tag{19}$$

where we have added a regularizing term $\left\| A \right\|_{HS}^2$. The solution is

$$\begin{aligned}
& \mathbb{E}_{XY} \left\| \varphi(y) - A\phi(x) \right\|_{\mathcal{G}}^2 + \lambda \left\| A \right\|_{HS}^2 \\
=& \mathbb{E}_Y L(Y,Y) - 2 \left\langle C_{YX}, A \right\rangle_{HS} + \left\langle A^*A, C_{XX} \right\rangle_{HS} + \lambda \left\langle A^*A, I \right\rangle \\
=& \mathbb{E}_Y L(Y,Y) - 2 \left\langle C_{YX}, A \right\rangle_{HS} + \left\langle A^*A, C_{XX} + \lambda I \right\rangle_{HS} \\
=& \mathbb{E}_Y L(Y,Y) - 2 \left\langle C_{YX}, B \left(C_{XX} + \lambda I \right)^{-1/2} \right\rangle_{HS} + \left\langle B, B \right\rangle_{HS} \qquad \text{defining } B := A \left(C_{XX} + \lambda I \right)^{1/2} \\
=& \mathbb{E}_Y L(Y,Y) - \left\| C_{YX} \left(C_{XX} + \lambda I \right)^{-1/2} \right\|_{HS}^2 + \left\| C_{YX} \left(C_{XX} + \lambda I \right)^{-1/2} - B \right\|_{HS}^2
\end{aligned}$$

where in the final row we complete the square. Since the final norm is non-negative, this is minimized when $B = C_{YX} \left(C_{XX} + \lambda I \right)^{-1/2}$, and hence $A = C_{YX} \left(C_{XX} + \lambda I \right)^{-1}$. This is the population expression for the regularized kernel conditional mean embedding (Song et al., 2009). To condition on a particular value of x, we apply this operator to its feature map $\phi(x) \in \mathcal{F}$.

Because this is the solution to a least squares regression problem, the regularization term λ can be found by cross-validation, replacing A with its empirical estimate, and the expectations in the loss with empirical expectations. Note also that density estimation is never required when computing conditional expectations in this manner.

D.2. The general vector-valued regression setting

We now describe a modified formulation of the problem of minimizing (18), where the mapping $\mu_{(x)}$ is directly from the set $\mathcal{X}$ to the Hilbert space $\mathcal{G}$. This might seem to be a minor change from the previous formulation (i.e., rather than having a map $\mu : \mathcal{F} \to \mathcal{G}$, we define a new map $\tilde{\mu} := \mu \circ \phi$ which is the composition of the feature map and the conditional mean embedding). In the latter formulation, however, it can be shown that the conditional mean embedding is a special case of (Micchelli & Pontil, 2005)[Thm. 4.1].

To compare directly with (Micchelli & Pontil, 2005), we move to the finite sample setting, replacing $\mathbb{E}_{Y \sim P(\cdot|X)}[f]$ with an estimate based on a sample $\{(x_i, y_i)\}_{i=1}^m$. At each x_i the sample estimate of $\mathbb{E}_{Y \sim P(\cdot|x_i)}[f]$ is just $f(y_i)$, and the loss function we minimize is $\sum_{i=1}^m ||L(y_i, \cdot) - \tilde{\mu}_{(x_i)}||_L^2$. We add a regularizer to make the problem well-posed,

$$\hat{\mu} = \operatorname*{argmin}_{\tilde{\mu}} \left\{ \sum_{i=1}^m ||L(y_i, \cdot) - \tilde{\mu}_{(x_i)}||_L^2 + \lambda ||\tilde{\mu}||_{\widetilde{K}}^2 \right\}. \tag{20}$$

Here $\widetilde{K}$ is the kernel for a $\mathcal{G}$-valued RKHS $\mathcal{H}_{\widetilde{K}} \subseteq \mathcal{G}^{\mathcal{X}}$: in other words, we have regression problem with the training data $\{(x_i, L(y_i, \cdot))\}_{i=1}^m$.

We now need to choose the kernel $\widetilde{K}$, which in the setting (Micchelli & Pontil, 2005) is a map $\mathcal{G} \to \mathcal{G}$ indexed by a pair $(x, x') \in \mathcal{X} \times \mathcal{X}$. Starting with the real-valued kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$, a natural choice for the RKHS $\mathcal{H}_{\widetilde{K}}$ has the inner product

$$\left\langle gK(x, \cdot), hK(x', \cdot) \right\rangle_{\widetilde{K}} := \left\langle g, h \right\rangle_L K(x, x') \tag{21}$$

for all $g, h \in \mathcal{H}_L$. Its easy to check that this satisfies the conditions to be a vector-valued RKHS in (Micchelli & Pontil, 2005) – in fact, it corresponds to the choice $\widetilde{K}(x, x') = K(x, x')\text{Id}$, where $\text{Id} : \mathcal{G} \to \mathcal{G}$ is the identity map on $\mathcal{G}$. We then recover the conditional mean embedding via (Micchelli & Pontil, 2005)[Thm. 4.1].