# Conditional mean embeddings as regressors

**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
**Sam Patterson**[*] SAM.X.PATTERSON@GMAIL.COM
**Arthur Gretton**[*,†] ARTHUR.GRETTON@GMAIL.COM
**Massimiliano Pontil** M.PONTIL@CS.UCL.AC.UK
CSML and [*]Gatsby Unit, University College London, UK, [†] MPI for Intelligent Systems

## Abstract

We demonstrate an equivalence between reproducing kernel Hilbert space (RKHS) embeddings of conditional distributions and vector-valued regressors. This connection introduces a natural regularized loss function which the RKHS embeddings minimise, providing an intuitive understanding of the embeddings and a justification for their use. Furthermore, the equivalence allows the application of vector-valued regression methods and results to the problem of learning conditional distributions. Using this link we derive a sparse version of the embedding by considering alternative formulations. Further, by applying convergence results for vector-valued regression to the embedding problem we derive minimax convergence rates which are $O(\log(n)/n)$ – compared to current state of the art rates of $O(n^{-1/4})$ – and are valid under milder and more intuitive assumptions. These minimax upper rates coincide with lower rates up to a logarithmic factor, showing that the embedding method achieves nearly optimal rates. We study our sparse embedding algorithm in a reinforcement learning task where the algorithm shows significant improvement in sparsity over an incomplete Cholesky decomposition.

## 1. Introduction/Motivation

In recent years a framework for embedding probability distributions into reproducing kernel Hilbert spaces (RKHS) has become increasingly popular (Smola et al., 2007). One example of this theme has been the representation of conditional expectation operators as RKHS functions, known as *conditional mean embeddings* (Song et al., 2009). Conditional expectations appear naturally in many machine learning tasks, and the RKHS representation of such expectations has two important advantages: first, conditional mean embeddings do not require solving difficult intermediate problems such as density estimation and numerical integration; and second, these embeddings may be used to compute conditional expectations directly on the basis of observed samples. Conditional mean embeddings have been successfully applied to inference in graphical models, reinforcement learning, subspace selection, and conditional independence testing (Fukumizu et al., 2008; 2009; Song et al., 2009; 2010; Grünewälder et al., 2012).

The main motivation for conditional means in Hilbert spaces has been to generalize the notion of conditional expectations from finite cases (multivariate Gaussians, conditional probability tables, and so on). Results have been established for the convergence of these embeddings in RKHS norm (Song et al., 2009; 2010), which show that conditional mean embeddings behave in the way we would hope (i.e., they may be used in obtaining conditional expectations as inner products in feature space, and these estimates are consistent under smoothness conditions). Despite these valuable results, the characterization of conditional mean embeddings remains incomplete, since these embeddings have not been defined in terms of the optimizer of a given *loss function*. This makes it difficult to extend these results, and has hindered the use of standard techniques like cross-validation for parameter estimation.

In this paper, we demonstrate that the conditional mean embedding is the solution of a vector-valued regression problem with a natural loss, resembling the standard Tikhonov regularized least-squares problem in multiple dimensions. Through this link, it is possible to access the rich the-

[1] **Equal contribution.** – Supplementary on arXiv.

Appearing in *Proceedings of the 29th International Conference on Machine Learning*, Edinburgh, Scotland, UK, 2012.

ory of vector-valued regression (Micchelli & Pontil, 2005; Carmeli et al., 2006; Caponnetto & De Vito, 2007; Caponnetto et al., 2008). We demonstrate the utility of this connection by providing novel characterizations of conditional mean embeddings, with important theoretical and practical implications. On the theoretical side, we establish novel convergence results for RKHS embeddings, giving a significant improvement over the rate of $O(n^{-1/4})$ due to Song et al. (2009; 2010). We derive a faster $O(\log(n)/n)$ rate which holds over large classes of probability measures, and requires milder and more intuitive assumptions. We also show our rates are optimal up to a $\log(n)$ term, following the analysis of Caponnetto & De Vito (2007). On the practical side, we derive an alternative sparse version of the embeddings which resembles the Lasso method, and provide a cross-validation scheme for parameter selection.

## 2. Background

In this section, we recall some background results concerning RKHS embeddings and vector-valued RKHS. For an introduction to scalar-valued RKHS we refer the reader to (Berlinet & Thomas-Agnan, 2004).

### 2.1. Conditional mean embeddings

Given sets $\mathcal{X}$ and $\mathcal{Y}$, with a distribution $P$ over random variables $(X, Y)$ from $\mathcal{X} \times \mathcal{Y}$ we consider the problem of learning expectation operators corresponding to the conditional distributions $P(Y|X = x)$ on $\mathcal{Y}$ after conditioning on $x \in \mathcal{X}$. Specifically, we begin with a kernel $L : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$, with corresponding RKHS $\mathcal{H}_L \subseteq \mathbb{R}^{\mathcal{Y}}$, and study the problem of learning, for every $x \in \mathcal{X}$, the *conditional expectation mapping* $\mathcal{H}_L \ni h \mapsto \mathbb{E}[h(Y)|X = x]$. Each such map can be represented as

$$\mathbb{E}[h(Y)|X = x] = \langle h, \mu(x)\rangle_L,$$

where the element $\mu(x) \in \mathcal{H}_L$ is called the *(conditional) mean embedding* of $P(Y|X = x)$. Note that, for every $x$, $\mu(x)$ is a function on $\mathcal{Y}$. It is thus apparent that $\mu$ is a mapping from $\mathcal{X}$ to $\mathcal{H}_L$, a point which we will expand upon shortly.

We are interested in the problem of estimating the embeddings $\mu(x)$ given an i.i.d. sample $\{(x_i, y_i)\}_{i=1}^n$ drawn from $P^n$. Following (Song et al., 2009; 2010), we define a second kernel $K : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ with associated RKHS $\mathcal{H}_K$, and consider the estimate

$$\hat{\mu}(x) := \sum_{i=1}^{n} \alpha_i(x) L(y_i, \cdot), \tag{1}$$

where $\alpha_i(x) = \sum_{j=1}^n W_{ij} K(x_j, x)$, and where $\boldsymbol{W} := (\boldsymbol{K} + \lambda n \boldsymbol{I})^{-1}$, $\boldsymbol{K} = (K(x_i, x_j))_{ij=1}^n$, and $\lambda$ is a chosen regularization parameter. This expression suggests that the conditional mean embedding is the solution to an underlying regression problem: we will formalize this link in Section 3. In the remainder of the present section, we introduce the necessary terminology and theory for vector valued regression in RHKSs.

### 2.2. Vector-valued regression and RKHSs

We recall some background on learning vector-valued functions using kernel methods (see Micchelli & Pontil, 2005, for more detail). We are given a sample $\{(x_i, v_i)\}_{i \le m}$ drawn i.i.d. from some distribution over $\mathcal{X} \times \mathcal{V}$, where $\mathcal{X}$ is a non-empty set and $(\mathcal{V}, \langle \cdot, \cdot \rangle_{\mathcal{V}})$ is a Hilbert space. Our goal is to find a function $f : \mathcal{X} \to \mathcal{V}$ with low error, as measured by

$$\mathbb{E}_{(X,V)}[||f(X) - V||_{\mathcal{V}}^2]. \tag{2}$$

This is the *vector-valued regression* problem (square loss).

One approach to the vector-valued regression problem is to model the regression function as being in a vector-valued RKHS of functions taking values in $\mathcal{V}$, which can be defined by analogy with the scalar valued case.

**Definition** A Hilbert space $(\mathcal{H}, \langle \cdot, \cdot \rangle_\Gamma)$ of functions $h : \mathcal{X} \to \mathcal{V}$ is an RKHS if for all $x \in \mathcal{X}, v \in \mathcal{V}$ the linear functional $h \mapsto \langle v, h(x)\rangle_{\mathcal{V}}$ is continuous.

The reproducing property for vector-valued RHKSs follows from this definition (see Micchelli & Pontil, 2005, Sect. 2). By the Riesz representation theorem, for each $x \in \mathcal{X}$ and $v \in \mathcal{V}$, there exists a linear operator from $\mathcal{V}$ to $\mathcal{H}_\Gamma$ written $\Gamma_x v \in \mathcal{H}_\Gamma$, such that for all $h \in \mathcal{H}_\Gamma$,

$$\langle v, h(x)\rangle_{\mathcal{V}} = \langle h, \Gamma_x v\rangle_\Gamma.$$

It is instructive to compare to the scalar-valued RKHS $\mathcal{H}_K$, for which the linear operator of evaluation $\delta_x$ mapping $h \in \mathcal{H}_K$ to $h(x) \in \mathbb{R}$ is continuous: then Riesz implies there exists a $K_x$ such that $yh(x) = \langle h, yK_x\rangle_K$.

We next introduce the vector-valued reproducing kernel, and show its relation to $\Gamma_x$. Writing as $\mathcal{L}(\mathcal{V})$ the space of bounded linear operators from $\mathcal{V}$ to $\mathcal{V}$, the reproducing kernel $\Gamma(x, x') \in \mathcal{L}(\mathcal{V})$ is defined as

$$\Gamma(x, x')v = (\Gamma_{x'} v)(x) \in \mathcal{V}.$$

From this definition and the reproducing property, the following holds (Micchelli & Pontil, 2005, Prop. 2.1).

**Proposition 2.1.** *A function $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{V})$ is a kernel if it satisfies: (i) $\Gamma(x, x') = \Gamma(x', x)^*$, (ii) for all $n \in \mathbb{N}, \{(x_i, v_i)\}_{i \le n} \subseteq \mathcal{X} \times \mathcal{V}$ we have that* $\sum_{i,j \le n} \langle v_i, \Gamma(x_i, x_j) v_j\rangle_{\mathcal{V}} \ge 0$.

It is again helpful to consider the scalar case: here, $\langle K_x, K_{x'}\rangle_K = K(x, x')$, and to every positive definite kernel $K(x, x')$ there corresponds a unique (up to isometry) RKHS for which $K$ is the reproducing kernel. Similarly, if $\Gamma : \mathcal{X} \times \mathcal{X} \to \mathcal{L}(\mathcal{V})$ is a kernel in the sense of Proposition 2.1, there exists a unique (up to isometry) RKHS, with $\Gamma$ as its reproducing kernel (Micchelli & Pontil, 2005, Th. 2.1). Furthermore, the RKHS $\mathcal{H}_\Gamma$ can be described as the RKHS limit of finite sums; that is, $\mathcal{H}_\Gamma$ is up to isometry equal to the closure of the linear span of the set $\{\Gamma_x v : x \in \mathcal{X}, v \in \mathcal{V}\}$, wrt the RKHS norm $\| \cdot \|_\Gamma$.

Importantly, it is possible to perform regression in this setting. One approach to the vector-valued regression problem is to replace the unknown true error (2) with a sample-based estimate $\sum_{i=1}^n ||v_i - f(x_i)||_\mathcal{V}^2$, restricting $f$ to be an element of an RKHS $\mathcal{H}_\Gamma$ (of vector-valued functions), and regularizing w.r.t. the $\mathcal{H}_\Gamma$ norm, to prevent overfitting. We thus arrive at the following regularized empirical risk,

$$\widehat{\mathcal{E}}_\lambda(f) := \sum_{i=1}^n ||v_i - f(x_i)||_\mathcal{V}^2 + \lambda ||f||_\Gamma^2. \qquad (3)$$

**Theorem 2.2.** *(Micchelli & Pontil, 2005, Th. 4) If $f^*$ minimises $\widehat{\mathcal{E}}_\lambda$ in $\mathcal{H}_\Gamma$ then it is unique and has the form,*

$$f^* = \sum_{i=1}^n \Gamma_{x_i} c_i,$$

*where the coefficients $\{c_i\}_{i \le n}$, $c_i \in \mathcal{V}$ are the unique solution of the system of linear equations*

$$\sum_{i \le n} (\Gamma(x_j, x_i) + \lambda \delta_{ji}) c_i = v_j, \quad 1 \le j \le n.$$

In the scalar case we have that $|f(x)| \le \sqrt{K(x,x)} \|f\|_K$. Similarly it holds that $\|f(x)\| \le |||\Gamma(x,x)|||\ \|f\|_\Gamma$, where $|||\cdot|||$ denotes the operator norm (Micchelli & Pontil, 2005, Prop. 1). Hence, if $|||\Gamma(x,x)||| \le B$ for all $x$ then

$$\|f(x)\|_\mathcal{V} \le B \|f\|_\Gamma. \qquad (4)$$

Finally, we need a result that tells us when all functions in an RKHS are continuous. In the scalar case this is guaranteed if $K(x, \cdot)$ is continuous for all $x$ and $K$ is bounded. In our case we have (Carmeli et al., 2006)[Prop. 12]:

**Corollary 2.3.** *If $\mathcal{X}$ is a Polish space, $\mathcal{V}$ a separabe Hilbert space and the mapping $x \mapsto \Gamma(\cdot, x)$ is continuous, then $\mathcal{H}_\Gamma$ is a subset of the set of continuous functions from $\mathcal{X}$ to $\mathcal{V}$.*

## 3. Estimating conditional expectations

In this section, we show the problem of learning conditional mean embeddings can be naturally formalised in the framework of vector-valued regression, and in doing so we derive an equivalence between the conditional mean embeddings and a vector-valued regressor.

### 3.1. The equivalence between conditional mean embeddings and a vector-valued regressor

Conditional expectations $\mathbb{E}[h(Y)|X = x]$ are linear in the argument $h$ so that, when we consider $h \in \mathcal{H}_L$, the Riesz representation theorem implies the existence of an element $\mu(x) \in \mathcal{H}_L$ such that $\mathbb{E}[h(Y)|X = x] = \langle h, \mu(x) \rangle_L$ for all $h$. That being said, the dependence of $\mu$ on $x$ may be complicated. A natural optimisation problem associated to this approximation problem is to therefore find a function $\mu : \mathcal{X} \to \mathcal{H}_L$ such that the following objective is small

$$\mathcal{E}[\mu] := \sup_{\|h\|_L \le 1} \mathbb{E}_X \left[ \left( \mathbb{E}_Y[h(Y)|X] - \langle h, \mu(X) \rangle_L \right)^2 \right]. \quad (5)$$

Note that the risk function cannot be used directly for estimation, because we do not observe $\mathbb{E}_Y[h(Y)|X]$, but rather pairs $(X, Y)$ drawn from $P$. However, we can bound this risk function with a surrogate risk function that has a sample based version,

$$\begin{aligned} &\sup_{\|h\|_L \le 1} \mathbb{E}_X \left[ \left( \mathbb{E}_Y[h(Y)|X] - \langle h, \mu(X) \rangle_L \right)^2 \right] \\ &= \sup_{\|h\|_L \le 1} \mathbb{E}_X \left[ \left( \mathbb{E}_Y[\langle h, L(Y, \cdot) \rangle_L | X] - \langle h, \mu(X) \rangle_L \right)^2 \right] \\ &\le \sup_{\|h\|_L \le 1} \mathbb{E}_{X,Y} \left[ \langle h, L(Y, \cdot) - \mu(X) \rangle_L^2 \right] \\ &\le \sup_{\|h\|_L \le 1} \|h\|_L^2 \, \mathbb{E}_{X,Y} \left[ \|L(Y, \cdot) - \mu(X)\|_L^2 \right] \\ &= \mathbb{E}_{(X,Y)} \left[ ||L(Y, \cdot) - \mu(X)||_L^2 \right], \end{aligned} \qquad (6)$$

where the first and second bounds follow by Jensen's and Cauchy-Schwarz's inequalities, respectively. Let us denote this surrogate risk function as

$$\mathcal{E}_s[\mu] := \mathbb{E}_{(X,Y)} \left[ ||L(Y, \cdot) - \mu(X)||_L^2 \right]. \qquad (7)$$

The two risk functions $\mathcal{E}$ and $\mathcal{E}_s$ are closely related and in Section 3.3 we examine their relation.

We now replace the expectation in (6) with an empirical estimate, to obtain the sample-based loss,

$$\widehat{\mathcal{E}}_n[\mu] := \sum_{i=1}^n ||L(y_i, \cdot) - \mu(x_i)||_L^2. \qquad (8)$$

Taking (8) as our empirical loss, then following Section 2.2 we add a regularization term to provide a well-posed problem and prevent overfitting,

$$\widehat{\mathcal{E}}_{\lambda,n}[\mu] := \sum_{i=1}^n ||L(y_i, \cdot) - \mu(x_i)||_L^2 + \lambda ||\mu||_\Gamma^2. \quad (9)$$

We denote the minimizer of (9) by $\hat{\mu}_{\lambda,n}$,

$$\hat{\mu}_{\lambda,n} := \operatorname*{argmin}_\mu \left\{ \widehat{\mathcal{E}}_{\lambda,n}[\mu] \right\}. \qquad (10)$$

Thus, recalling (3), we can see that the problem (10) is posed as a vector-valued regression problem with the training data now considered as $\{(x_i, L(y_i, \cdot))\}_{i=1}^n$ (and we identify $\mathcal{H}_L$ with the general Hilbert space $\mathcal{V}$ of Section 2.2). From Theorem 2.2, the solution is

$$\hat{\mu}_{\lambda,n} = \sum_{i=1}^{n} \Gamma_{x_i} c_i, \tag{11}$$

where the coefficients $\{c_i\}_{i \le n}$, $c_i \in \mathcal{H}_L$ are the unique solution of the linear equations

$$\sum_{i \le n} (\Gamma(x_j, x_i) + \lambda \delta_{ji}) c_i = L(y_j, \cdot), \quad 1 \le j \le n.$$

It remains to choose the kernel $\Gamma$. Given a real-valued kernel $K$ on $\mathcal{X}$, a natural choice for the RKHS $\mathcal{H}_\Gamma$ would be the space of functions from $\mathcal{X}$ to $\mathcal{H}_L$ whose elements are defined as functions via $(h, K(x, \cdot))(x') := K(x, x')h$, which is isomorphic to $\mathcal{H}_L \otimes \mathcal{H}_K$, with inner product

$$\langle gK(x, \cdot), hK(x', \cdot) \rangle_\Gamma := \langle g, h \rangle_L K(x, x') \tag{12}$$

for all $g, h \in \mathcal{H}_L$. Its easy to check that this satisfies the conditions to be a vector-valued RKHS– in fact it corresponds to the choice $\Gamma(x, x') = K(x, x')\text{Id}$, where $\text{Id} : \mathcal{H}_L \to \mathcal{H}_L$ is the identity map on $\mathcal{H}_L$. The solution to (10) with this choice is then given by (11), with

$$\sum_{i \le n} (K(x_j, x_i) + \lambda \delta_{ji}) c_i = L(y_j, \cdot), \quad 1 \le j \le n$$

$$c_i = \sum_{j \le n} W_{ij} L(y_j, \cdot), \quad 1 \le i \le n,$$

where $\boldsymbol{W} = (\boldsymbol{K} + \lambda \boldsymbol{I})^{-1}$, which corresponds exactly the embeddings (1) presented in (Song et al., 2009; 2010) (after a rescaling of $\lambda$). Thus we have shown that the embeddings of Song et al. are the solution to a regression problem for a particular choice of operator-valued kernel. Further, the loss defined by (7) is a key error functional in this context since it is the objective which the estimated embeddings attempt to minimise. In Sec. 3.3 we will see that this does not always coincide with (5) which may be a more natural choice. In Sec. 4 we analyze the performance of the embeddings defined by (10) at minimizing the objective (7).

## 3.2. Some consequences of this equivalence

We derive some immediate benefits from the connection described above. Since the embedding problem has been identified as a regression problem with training set $\mathcal{D} := \{(x_i, L(y_i, \cdot))\}_{i=1}^m$, we can define a cross validation scheme for parameter selection in the usual way: by holding out a subsample $\mathcal{D}_{\text{val}} = \{(x_{t_i}, L(y_{t_j}, \cdot))\}_{j=1}^T \subset \mathcal{D}$, we can train embeddings $\hat{\mu}$ on $\mathcal{D} \backslash \mathcal{D}_{\text{val}}$ over a grid of kernel or regularization parameters, choosing the parameters achieving the best error $\sum_{j=1}^T ||\hat{\mu}(x_{t_j}) - L(y_{t_j}, \cdot)||_L^2$ on the validation set (or over many folds of cross validation). Another key benefit will be a much improved performance analysis of the embeddings, presented in Section 4.

| | |
|---|---|
| Input/Output space | (i) $\mathcal{X}$ is Polish.<br>(ii) $\mathcal{V}$ is separable. *(f.b.a.)*<br>(iii) $\exists C > 0$ such that $\forall x \in \mathcal{X}$<br>$\text{Tr}(\Gamma(x, x)) \le C$ holds. |
| Space of regressors | (iv) $\mathcal{H}_\Gamma$ is separable.<br>(v) All $\Gamma_x^*$ are HS. *(f.b.a.)*<br>(vi) $B : (x, y) \to \langle f, \Gamma(y, x) g \rangle_\mathcal{V}$<br>is measurable $\forall f, g \in \mathcal{V}$. |
| True distribution | (vii) $L(y, y) < \infty$ for all $y \in \mathcal{Y}$.<br>(viii) $\mathcal{E}_s[\mu^*] = \inf_{\mu \in \mathcal{H}_\Gamma} \mathcal{E}_s[\mu]$. |

*Table 1.* Assumptions for Corollary 4.1 and 4.2. f.b.a. stands for fulfilled by assumption that $\mathcal{V}$ is finite dimensional.

## 3.3. Relations between the error functionals $\mathcal{E}$ and $\mathcal{E}_s$

In Section 3.1 we introduced an alternative risk function $\mathcal{E}_s$ for $\mathcal{E}$, which we used to derive an estimation scheme to recover conditional mean embeddings. We now examine the relationship between the two risk functionals. When the true conditional expectation on functions $h \in \mathcal{H}_L$ can be represented through an element $\mu^* \in \mathcal{H}_\Gamma$ then $\mu^*$ minimises both objectives.

**Theorem 3.1** (Proof in App. A)**.** *If there exists a $\mu^* \in \mathcal{H}_\Gamma$ such that for any $h \in \mathcal{H}_L$: $\mathbb{E}[h|X] = \langle h, \mu^*(X) \rangle_L$ $P_\mathcal{X}$-a.s., then $\mu^*$ is the $P_\mathcal{X}$-a.s. unique minimiser of both objectives:*

$$\mu^* = \operatorname*{argmin}_{\mu \in \mathcal{H}_\Gamma} \mathcal{E}[\mu] = \operatorname*{argmin}_{\mu \in \mathcal{H}_\Gamma} \mathcal{E}_s[\mu] \quad P_\mathcal{X} \; a.s.$$

Thus in this case, the embeddings of Song et al. (e.g. 2009) minimise both (5) and (7). More generally, however, this may not be the case. Let us define an element $\hat{\mu}$ that is $\delta$ close w.r.t. the error $\mathcal{E}_s$ to the minimizer $\mu'$ of $\mathcal{E}_s$ in $\mathcal{H}_\Gamma$ (this might for instance be the minimizer of the empirical regularized loss for sufficiently many samples). We are interested in finding conditions under which $\mathcal{E}(\hat{\mu})$ is not much worse than a good *approximation* $\mu^*$ in $\mathcal{H}_\Gamma$ to the conditional expectation. The sense in which $\mu^*$ approximates the conditional expectation is somewhat subtle: $\mu^*$ must closely approximate the conditional expectation of functions $\mu \in \mathcal{H}_\Gamma$ under the original loss $\mathcal{E}$ (note that the loss $\mathcal{E}$ was originally defined in terms of functions $h \in \mathcal{H}_L$).

**Theorem 3.2** (Proof in App. A)**.** *Let $\mu'$ be a minimiser of $\mathcal{E}_s$ and $\hat{\mu}$ be an element of $\mathcal{H}_\Gamma$ with $\mathcal{E}_s[\hat{\mu}] \le \mathcal{E}_s[\mu'] + \delta$. Define, $\mathcal{A} := \{(\eta, \tilde{\mu}) \,|\, \eta^2 =$*

$\sup_{||\mu||_\Gamma \leq 1} \mathbb{E}_X \left[\mathbb{E}_Y[\mu(X)|X] - \langle \mu(X), \tilde{\mu}(X) \rangle_L\right]^2\}$ , *then*

$$\mathcal{E}[\hat{\mu}] \leq \inf_{(\eta,\mu^*)\in\mathcal{A}} \left(\sqrt{\mathcal{E}[\mu^*]} + \sqrt{8\eta(\|\mu^*\|_\Gamma + \|\hat{\mu}\|_\Gamma)} + \delta^{\frac{1}{2}}\right)^2 .$$

Apart from the more obvious condition that $\delta$ be small, the above theorem suggests that $||\hat{\mu}||_\Gamma$ should also be made small for the solution $\hat{\mu}$ to have low error $\mathcal{E}$. In other words, even in the infinite sample case, the regularization of $\hat{\mu}$ in $\mathcal{H}_\Gamma$ is important.

## 4. Better convergence rates for embeddings

The interpretation of the mean embedding as a vector valued regression problem allows us to apply regression minimax theorems to study convergence rates of the embedding estimator. These rates are considerably better than the current state of the art for the embeddings, and hold under milder and more intuitive assumptions.

We start by comparing the statements which we derive from (Caponnetto & De Vito, 2007, Thm.s 1 and 2) with the known convergence results for the embedding estimator. We follow this up with a discussion of the rates and a comparison of the assumptions.

### 4.1. Convergence theorems

We address the performance of the embeddings defined by (10) in terms of asymptotic guarantees on the loss $\mathcal{E}_s$ defined by (7). Caponnetto & De Vito (2007) study uniform convergence rates for regression. Convergence rates of learning algorithms can not be uniform on the set of all probability distributions if the output vector space is an infinite dimensional RKHS (Caponnetto & De Vito, 2007)[p. 4]. It is therefore necessary to restrict ourselves to a subset of probability measures. This is done by Caponnetto & De Vito (2007) by defining families of probability measures $\mathscr{P}(b,c)$ indexed by two parameters $b \in ]1,\infty]$ and $c \in [1,2]$. We discuss the family $\mathscr{P}(b,c)$ in detail below. The important point at the moment is that $b$ and $c$ affect the optimal schedule for the regulariser $\lambda$ and the convergence rate. The rate of convergence is better for higher $b$ and $c$ values. Caponnetto & De Vito (2007) provide convergence rates for all choices of $b$ and $c$. We restrict ourself to the best case $b = \infty, c > 1$ and the worst case[1] $b = 2, c = 1$.

We recall that the estimated conditional mean embeddings $\hat{\mu}_{\lambda,n}$ are given by (10), where $\lambda$ is a chosen regularization parameter. We assume $\lambda_n$ is chosen to follow a specific schedule, dependent upon $n$: we denote by $\hat{\mu}_n$ the embeddings following this schedule and $\mu' := \operatorname{argmin}_{\mu\in\mathcal{H}_\Gamma} \mathcal{E}_s[\mu]$. Given this optimal rate of decrease for $\lambda_n$, Thm. 1 of Caponnetto & De Vito (2007) yields the following convergence statements for the estimated embeddings, under assumptions to be discussed in Section 4.2.

**Corollary 4.1.** *Let $b = \infty, c > 1$ then for every $\epsilon > 0$ there exists a constant $\tau$ such that*

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}(b,c)} P^n \left[\mathcal{E}_s[\hat{\mu}_n] - \mathcal{E}_s[\mu'] > \tau\frac{1}{n}\right] < \epsilon.$$

*Let $b = 2$ and $c = 1$ then for every $\epsilon > 0$ there exists a constant $\tau$ such that*

$$\limsup_{n\to\infty} \sup_{P\in\mathscr{P}(b,c)} P^n \left[\mathcal{E}_s[\hat{\mu}_n] - \mathcal{E}_s[\mu'] > \tau\left(\frac{\log n}{n}\right)^{\frac{2}{3}}\right] < \epsilon.$$

The rate for the estimate $\hat{\mu}_n$ can be complemented with minimax lower rates for vector valued regression (Caponnetto & De Vito, 2007)[Th. 2] in the case that $b < \infty$.

**Corollary 4.2.** *Let $b = 2$ and $c = 1$ and let $\Lambda_n := \{l_n \,|l_n : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}_\Gamma\}$ be the set of all learning algorithm working on $n$ samples, outputting $\nu_n \in \mathcal{H}_\Gamma$. Then for every $\epsilon > 0$ there exists a constant $\tau > 0$ such that*

$$\liminf_{n\to\infty} \inf_{l_n\in\Lambda_n} \sup_{P\in\mathscr{P}(b,c)} P^n \left[\mathcal{E}_s[\nu_n] - \mathcal{E}_s[\mu'] > \tau\left(\frac{1}{n}\right)^{\frac{2}{3}}\right] > 1-\epsilon.$$

This corollary tells us that there exists no learning algorithm which can achieve better rates than $n^{-\frac{2}{3}}$ uniformly over $\mathscr{P}(2,1)$, and hence the estimate $\hat{\mu}_n$ is optimal up to a logarithmic factor.

**State of the art results for the embedding** The current convergence result for the embedding is proved by Song et al. (2010, Th.1). A crucial assumption that we discuss in detail below is that the mapping $x \mapsto \mathbb{E}[h(Y)|X = x]$ is in the RKHS $\mathcal{H}_K$ of the real valued kernel, i.e. that for all $h \in \mathcal{H}_L$ we have that there exists a $f_h \in \mathcal{H}_K$, such that

$$\mathbb{E}[h(Y)|X = x] = f_h(x). \tag{13}$$

The result of Song et al. implies the following (see App. C): if $K(x,x) < B$ for all $x \in \mathcal{X}$ and the schedule $\lambda(n) = n^{-1/4}$ is used: for a fixed probability measure $P$, there exists a constant $\tau$ such that

$$\lim_{n\to\infty} P^n \left[\mathcal{E}[\hat{\mu}_n] > \tau\left(\frac{1}{n}\right)^{\frac{1}{4}}\right] = 0, \tag{14}$$

where $\hat{\mu}_n(x)$ is the estimate from Song et al. No complementary lower bounds were known until now.

[1] Strictly speaking the worst case is $b \downarrow 1$ (see supp.).

**Comparison** The first thing to note is that under the assumption that $\mathbb{E}[h|X]$ is in the RKHS $\mathcal{H}_K$ the minimiser of $\mathcal{E}_s$ and $\mathcal{E}$ are a.e. equivalent due to Theorem 3.1: the assumption implies a $\mu^* \in \mathcal{H}_\Gamma$ exists with $\mathbb{E}[h|X] = \langle h, \mu^*(X)\rangle_L$ for all $h \in \mathcal{V}$ (see App. B.4 for details). Hence, under this assumption, the statements from eq. 14 and Cor. 4.1 ensure we converge to the true conditional expectation, and achieve an error of 0 in the risk $\mathcal{E}$.

In the case that this assumption is not fullfilled and eq. 14 is not applicable, Cor. 4.1 still tells us that we converge to the minimiser of $\mathcal{E}_s$. Coupling this statement with Thm. 3.2 allows us to bound the distance to the minimal error $\mathcal{E}[\mu^*]$, where $\mu^* \in \mathcal{H}_\Gamma$ minimises $\mathcal{E}$.

The other main differences are obviously the rates, and that Cor. 4.1 bounds the error uniformly over a space of probability measures, while eq. 14 provides only a point-wise statement (i.e., for a fixed probability measure $P$).

### 4.2. Assumptions

**Cor. 4.1 and 4.2** Our main assumption is that $\mathcal{H}_L$ is finite dimensional. It is likely that this assumption can be weakened, but this requires a deeper analysis.

The assumptions of Caponnetto & De Vito (2007) are summarized in Table 1, where we provide details in App. B.2. App. B.1 contains simple and complete assumptions that ensure all statements in the table hold. Beside some measure theoretic issues, the assumptions are fulfilled if for example, 1) $\mathcal{X}$ is a compact subset of $\mathbb{R}^n$, $\mathcal{Y}$ is compact, $\mathcal{H}_L$ is a finite dimensional RKHS, $\Gamma$ and $L$ are continuous; 2) $\mathcal{E}_s[\mu'] = \inf_{\mu \in \mathcal{H}_\Gamma} \mathcal{E}_s[\mu]$. This last condition is unintuitive, but can be rewritten in the following way:

**Theorem 4.3** (Proof in App.)**.** *Let $||h||_\mathcal{V}, ||\mu(x) - h||_\mathcal{V}$ be integrable for all $h \in \mathcal{H}_\Gamma$ and let $\mathcal{V}$ be finite dimensional. Then there exists a $\mu' \in \mathcal{H}_\Gamma$ with $\mathcal{E}_s[\mu'] = \inf_{\mu \in \mathcal{H}_\Gamma} \mathcal{E}_s[\mu]$ iff a $B > 0$ exists and a sequence $\{\mu_n\}_{n\in\mathbb{N}}$ with $\mathcal{E}_s[\mu_n] \leq \inf_{\mu \in \mathcal{H}_\Gamma} \mathcal{E}_s[\mu] + 1/n$ and $||\mu_n||_\Gamma < B$.*

The intuition is that the condition is not fulfilled if we need to make $\mu_n$ more and more complex (in the sense of a high RKHS norm) to optimize the risk.

**Definition and discussion of $\mathscr{P}(b, c)$** The family of probability measures $\mathscr{P}(b, c)$ is characterized through spectral properties of the kernel function $\Gamma$. The assumptions correspond to assumptions on the eigenvalues in Mercer's theorem in the real valued case, i.e. that there are finitely many eigenvalues or that the eigenvalues decrease with a certain rate. In detail, define the operator $A$ through $A(\phi)(x') := \int_\mathcal{X} \Gamma(x', x)\phi(x) dP_\mathcal{X}$, where $\phi \in L^2(P_\mathcal{X})$. $A$ can be written as (Caponnetto & De Vito, 2007, Rem. 2) $A = \sum_{n=1}^N \lambda_n \langle \cdot, \phi_n\rangle_P \phi_n$, where the inner product is the $L^2$ inner product with measure $P_\mathcal{X}$ and $N = \infty$ is allowed. As in the real valued case, the eigendecomposition depends on the measure on the space $\mathcal{X}$ but is independent of the distribution on $\mathcal{Y}$. The eigendecomposition measures the complexity of the kernel, where the lowest complexity is achieved for finite $N$ — that is, the case $b = \infty, c > 1$ — and has highest complexity if the eigenvalues decrease with the slowest possible rate, $\lambda_n < C/n$ for a constant $C$. The case $b = 2, c = 1$ correspond to a slightly faster decay, namely, $\lambda_n < C/n^2$. In essence, there are no assumptions on the distribution on $\mathcal{Y}$, but only on the complexity of the kernel $\Gamma$ as measured with $P_\mathcal{X}$.

**Embedding** The results of Song et al. (2010) do not rely on the assumption that $\mathcal{V}$ is finite dimensional. Other conditions on the distribution are required, however, which are challenging to verify. To describe these conditions, we recall the real-valued RKHS $\mathcal{H}_K$ with kernel $K$, and define the uncentred cross-covariance operator $C_{YX} : \mathcal{H}_K \to \mathcal{H}_L$ such that $\langle g, C_{YX} f\rangle_{\mathcal{H}_L} = \mathbb{E}_{XY}(f(X)g(Y))$, with the covariance operator $C_{XX}$ defined by analogy. One of the two main assumptions of Song et al. is that $C_{YX}C_{XX}^{-3/2}$ needs to be Hilbert-Schmidt. The covariances $C_{YX}$ and $C_{XX}$ are compact operators, meaning $C_{XX}$ is not invertible when $\mathcal{H}_K$ is infinite dimensional (this gives rise to a notational issue, although the "product" operator $C_{YX}C_{XX}^{-3/2}$ may still be defined). Whether $C_{YX}C_{XX}^{-3/2}$ is Hilbert-Schmidt (or even bounded) will depend on the underlying distribution $P_{XY}$ and on the kernels $K$ and $L$. At this point, however, there is no easy way to translate properties of $P_{XY}$ to guarantees that the assumption holds.

The second main assumption is that the conditional expectation can be represented as an RKHS element (see App B.4). Even for rich RKHSs (such as universal RKHSs), it can be challenging to determine the associated conditions on the distribution $P_{XY}$. For simple finite dimensional RKHSs, the assumption may fail, as shown below.

**Corollary 4.4** (Proof in App. C )**.** *Let $\mathcal{V}$ be finite dimensional such that a function $\tilde{h} \in \mathcal{V}$ exists with $\tilde{h}(y) \geq \epsilon > 0$ for all $y \in \mathcal{Y}$. Furthermore, let $\mathcal{X} := [-1, 1]$ and the reproducing kernel for $\mathcal{H}_K$ be $K(x, y) = xy$. Then there exists no measure for which the assumption from eq. (13) can be fulfilled.*

## 5. Sparse embeddings

In many practical situations it is desirable to approximate the conditional mean embedding by a sparse version which involves a smaller number of parameters. For example, in the context of reinforcement learning and planning, the sample size $n$ is large and we want to use the embeddings over and over again, possibly on many different tasks and over a long time frame.

Here we present a technique to achieve a sparse approximation of the sample mean embedding. Recall that this is given by the formula (cf. equation (11))

$$\hat{\mu}(x) = \sum_{i,j=1}^{n} W_{ij} K(x_i, x) L(y_j, \cdot),$$

where $W = (K + n\lambda I)^{-1}$. A natural approach to find a sparse approximation of $\hat{\mu}$ is to look for a function $\mu$ which is close to $\hat{\mu}$ according to the RKHS norm $\| \cdot \|_\Gamma$ (in App. D we establish a link between this objective and our cost function $\mathcal{E}$). In the special case that $\Gamma = KId$ this amounts to solving the optimization problem

$$\min_{M \in \mathbb{R}^{n \times n}} f(M - W) + \gamma \|M\|_{1,1} \tag{15}$$

where $\gamma$ is a positive parameter, $\|W\|_{1,1} := \sum_{i,j}^{n} |M_{ij}|$ and

$$f(M) = \left\| \sum_{i,j=1}^{n} M_{ij} K(x_i, \cdot) L(y_j, \cdot) \right\|_{K \otimes L} . \tag{16}$$

Problem (15) is equivalent to a kind of Lasso problem with $n^2$ variables: when $\gamma = 0$, $M = W$ at the optimum and the approximation error is zero, however as $\gamma$ increases, the approximation error increases as well, but the solution obtained becomes sparse (many of the elements of matrix $M$ are equal to zero).

A direct computation yields that the above optimization problem is equivalent to

$$\min_{M \in \mathbb{R}^{n \times n}} \operatorname{tr}((M-W)^\top K (M-W) L) + \gamma \sum_{i,j=1}^{n} |M_{ij}|. \tag{17}$$

In the experiments in Section 5.1, we solve problem (17) with FISTA (Beck & Teboulle, 2009), an optimal first order method which requires $O(1/\sqrt{\epsilon})$ iterations to reach a $\epsilon$ accuracy of the minimum value in (17), with a cost per iteration of $O(n^2)$ in our case. The algorithm is outlined below, where $S_\gamma(Z_{ij}) = \operatorname{sign}(Z_{ij})(|Z_{ij}| - \gamma)_+$ and $(z)_+ = z$ if $z > 0$ and zero otherwise.

**Algorithm 1** LASSO-like Algorithm

```
input: $W, \gamma, K, L$ output: $M$
$Z_1 = Q_1 = 0, \theta_1 = 1, C = \|K\| \, \|L\|$
for t=1,2,... do
    $Z_{t+1} = S_{\gamma C}\left(Q_t - C\,(K\,Q_t\,L - G\,W\,L)\right)$
    $\theta_{t+1} = \frac{1+\sqrt{1+4\theta_t^2}}{2}$
    $Q_{t+1} = Z_{t+1} + \frac{\theta_t - 1}{\theta_{t+1}}(Z_t - Z_{t+1})$
end for
```

Other sparsity methods could also be employed. For example, we may replace the norm $\| \cdot \|_{1,1}$ by a block $\ell_1/\ell_2$ norm. That is, we may choose the norm $\|M\|_{2,1} := \sum_{j=1}^{n} \sqrt{\sum_{i=1}^{n} M_{ij}^2}$, which is the sum of the $\ell_2$ norms of the rows of $M$. This penalty function encourages sparse approximations which use few input points but all the outputs. Similarly, the penalty $\|M^\top\|_{2,1}$ will sparsify over the outputs. Finaly, if we wish to remove many pair of examples we may use the more sophisticated penalty $\sum_{i,j=1}^{n} \sqrt{\sum_{k=1}^{m} M_{ik}^2 + M_{kj}^2}$.

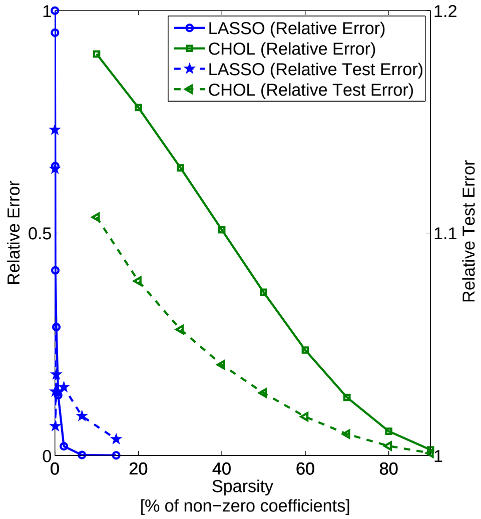

*Figure 1.* Comparison between our sparse algorithm and an incomplete Cholesky decomposition. The x-axis shows the level of sparsity, where on the right side the original solution is recovered. The y-axis shows the distance to the dense solution and the test error relative to the dense solution.

### 5.1. Experiment

We demonstrate here that the link between vector-valued regression and the mean embeddings can be leveraged to develop useful embedding alternatives that exploit properties of the regression formulation: we apply the sparse algorithm to a challenging reinforcement learning task. The sparse algorithm makes use of the labels, while other algorithms to sparsify the embeddings without our regression interpretation cannot make use of these. In particular, a popular method to sparsify is the incomplete Cholesky decomposition (Shawe-Taylor & Cristianini, 2004), which sparsifies based on the distribution on the input space $\mathcal{X}$ only. We compare to this method in the experiments.

The reinforcement learning task is the under-actuated pendulum swing-up problem from 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 $r(\theta, \omega) = \exp(-\theta^2 - 0.2\omega^2)$. The learning algorithm is a kernel method which uses the mean embedding estimator to perform policy iteration (Grünewälder et al., 2012). Sparse solutions are in this task very useful as the policy iteration applies the mean embedding many times to perform updates. The input space has $4$ dimensions (sine and cosine of the angle, angular velocity and applied torque), while the output has $3$ (sine and cosine of the angle and angular velocity). We sample uniformly a training set of $200$ examples and learn the mean conditional embedding using the direct method (1). We then compare the sparse approximation obtained by Algorithm 1 using different values of the parameter $\gamma$ to the approximation obtained via an incomplete Cholesky decomposition at different levels of sparsity. We assess the approximations using the test error and (16), which is an upper bound on the generalization error (see App. D) and report the results in Figure 1.

## 6. Outlook

We have established a link between vector-valued regression and conditional mean embeddings. On the basis of this link, we derived a sparse embedding algorithm, showed how cross-validation can be performed, established better convergence rates under milder assumptions, and complemented these upper rates with lower rates, showing that the embedding estimator achieves near optimal rates.

There are a number of interesting questions and problems which follow from our framework. It may be valuable to employ other kernels $\Gamma$ in place of the kernel $K(x,y)\mathrm{Id}$ that leads to the mean embedding, so as to exploit knowledge about the data generating process. As a related observation, for the kernel $\Gamma(x,y) := K(x,y)\mathrm{Id}$, $\Gamma_x$ is not a Hilbert-Schmidt operator if $\mathcal{V}$ is infinite dimensional, as

$$||\Gamma_x||^2_{HS} = K(x,x) \sum_{i=1}^{\infty} \langle e_i, \mathrm{Id}\, e_i \rangle_L = \infty,$$

however the convergence results from (Caponnetto & De Vito, 2007) assume $\Gamma_x$ to be Hilbert-Schmidt. While this might simply be a result of the technique used in (Caponnetto & De Vito, 2007), it might also indicate a deeper problem with the standard embedding estimator, namely that if $\mathcal{V}$ is infinite dimensional then the rates degrade. The latter case would have a natural interpretation as an overfitting effect, as $\mathrm{Id}$ does not "smooth" the element $h \in \mathcal{V}$.

Our sparsity approach can potentially be equipped with other regularisers that cut down on rows and columns of the $W$ matrix in parallel. Certainly, ours is not the only sparse regression approach, and other sparse regularizers might yield good performance on appropriate problems.

## Acknowledgements

The authors want to thank for the support of the EPSRC #EP/H017402/1 (CARDyAL) and the European Union #FP7-ICT-270327 (Complacs).

## References

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM Journal of Imaging Sciences*, 2(1):183–202, 2009.

Berlinet, A. and Thomas-Agnan, C. *Reproducing kernel Hilbert spaces in Probability and Statistics*. Kluwer, 2004.

Caponnetto, A. and De Vito, E. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7(3):331–368, 2007.

Caponnetto, A., Micchelli, C.A., Pontil, M., and Ying, Y. Universal multi-task kernels. *JMLR*, 9, 2008.

Carmeli, C., De Vito, E., and Toigo, A. Vector valued reproducing kernel Hilbert spaces of integrable functions and mercer theorem. *Analysis and Applications*, 4(4):377–408, 2006.

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

Fremlin, D.H. *Measure Theory Volume 1: The Irreducible Minimum*. Torres Fremlin, 2000.

Fremlin, D.H. *Measure Theory Volume 4: Topological Measure Spaces*. Torres Fremlin, 2003.

Fukumizu, K., Gretton, A., Sun, X., and Schölkopf, B. Kernel measures of conditional dependence. In *NIPS 20*, 2008.

Fukumizu, K., Bach, F., and Jordan, M. Kernel dimension reduction in regression. *The Annals of Statistics*, 37(4), 2009.

Fukumizu, K., Song, L., and Gretton, A. Kernel Bayes' rule. pp. 1737–1745, 2011.

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

Kallenberg, O. *Foundations of Modern Probability*. Springer, 2nd edition, 2001.

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

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

Smola, A., Gretton, A., Song, L., and Schölkopf, B. A Hilbert space embedding for distributions. In *ALT*. Springer, 2007.

Song, L., Huang, J., Smola, A. J., and Fukumizu, K. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *ICML*, pp. 121, 2009.

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

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

Werner, D. *Funktionalanalysis*. Springer, 4th edition, 2002.