Proceedings of the 23rd International Joint Conference on Artificial Intelligence (IJCAI), Beijing, 2013

# Path Integral Control by Reproducing Kernel Hilbert Space Embedding

**Konrad Rawlik**
School of Informatics
University of Edinburgh

**Marc Toussaint**
Inst. für Parallele und Verteilte Systeme
Universität Stuttgart

**Sethu Vijayakumar**
School of Informatics
University of Edinburgh

## Abstract

We present an embedding of stochastic optimal control problems, of the so called path integral form, into reproducing kernel Hilbert spaces. Using consistent, sample based estimates of the embedding leads to a model-free, non-parametric approach for calculation of an approximate solution to the control problem. This formulation admits a decomposition of the problem into an invariant and task dependent component. Consequently, we make much more efficient use of the sample data compared to previous sample based approaches in this domain, e.g., by allowing sample re-use across tasks. Numerical examples on test problems, which illustrate the sample efficiency, are provided.

## 1 Introduction

While solving general non-linear *stochastic optimal control* (SOC) and *Reinforcement Learning* (RL) problems remains challenging, some recent work by [Kappen, 2005] has identified a class of problems that admit closed form solutions. Although these solutions require evaluation of a path integral – equivalent to evaluation of a partition function, which in itself is a hard problem – they allow for the application of Monte Carlo and Variational methods, leading to several practical applications [Theodorou *et al.*, 2010; Broek *et al.*, 2011]. In the special case of linear dynamics and quadratic costs, the required path integral can be evaluated analytically based on linear operators acting on state vectors. Here, we show that, analogously, a suitable embedding of the path integral into a *reproducing kernel Hilbert space* (RKHS) allows its evaluation in terms of covariance operators acting on elements of the Hilbert space. While this in itself does not yield a tractable solution to the SOC problem, consistent estimators of the required operators give rise to efficient non-parametric algorithms.

The change of perspective from the direct estimation of the path integral (which previous applications of Monte Carlo methods aimed at) to estimation of operators allows to overcome several shortcomings of previous methods while maintaining many of their advantages. Most importantly, it can significantly reduce the sample complexity by splitting the problem appropriately into an invariant and task varying component, allowing efficient sample re-use across tasks and leading to a form of transfer learning – contrast this to the situation where any change in the task including, for e.g., different start states, necessitate acquiring new samples [Theodorou *et al.*, 2010; 2009]. Additionally, the approach remains model-free, allowing its application to the RL setting. This is in contrast to variational [Mensink *et al.*, 2010] or function approximation [Zhong and Todorov, 2011a; 2011b] approaches, from which it is further distinguished through convergence guarantees. The RKHS embedding make the operators state-dimensionality independent, leading to better scalability, while prior knowledge about both tasks and dynamics can be effectively incorporated by informing choices of sampling procedures and kernel.

It is worth noting that, while we choose to frame our approach in the context of path integral stochastic optimal control, it is not restricted to problems which fall into this class. The formalisms of linearly solvable MDPs by [Todorov, 2007], inference control by [Toussaint, 2009] and free energy control of [Friston *et al.*, 2010] all require solving an underlying problem of equivalent form, making the methods proposed directly applicable in these contexts. Furthermore [Rawlik *et al.*, 2012a] discusses a formulation which generalizes path integral control to derive an optimal policy for general SOC problems. Finally, while we focus on finite horizon problems, path integral formulations for discounted and average cost infinite horizon problems have been proposed by [Todorov, 2009], as well as by [Broek *et al.*, 2010] for risk sensitive control.

Before proceeding, let us briefly outline the structure of the paper and give an informal overview about the key concepts involved in our approach. In Section 2 we first briefly review the formulation of a SOC problem in the path integral control framework of [Kappen, 2005]. This formulation has two specific properties relevant for our approach. First, it yields a closed form solution for the optimal policy in terms of a state desirability function $\Psi$ (c.f. (3–4)). Second, this function $\Psi$ can be expressed as a conditional expectation of the product of an immediate cost and a future desirability (c.f. (7)). By identifying this expression with an inner product in a RKHS, we can compute the desirability function. We formalise this notion in Section 3.1, where we address the model-based case and derive the analytical form of the RKHS based evaluation (c.f. (13)). In Section 3.2, we consider the model-free case and discuss that this operator can be estimated from transition samples, leading to implementation in form of a finite dimensional inner product. To summarise, we solve the SOC problem by, (i) collecting transition samples and estimating the embedding operator, (ii) representing the desirability function and immediate cost as elements of RKHSs, (iii) recursively evaluating the desirability function as a series of inner products based on the estimated operator – n.b. this process involves a series of matrix multiplications – and finally (iv) computing the optimal controls based on the obtained desirability function. In Section 4, we build on this basic methodology and discuss a series of alternative estimators with either reduced computational costs or which allow for more efficient use of the sample data. Finally, in Section 5, we validate the proposed methods in experiments.

## 2 Path Integral Control

In this section we briefly review the path integral approach to stochastic optimal control as proposed by [Kappen, 2005] (see also [Kappen, 2011; Theodorou *et al.*, 2010]). Let $\mathbf{x} \in \mathbb{R}^{d_x}$ be the system state and $\mathbf{u} \in \mathbb{R}^{d_u}$ the control signals. Consider a continuous time stochastic system of the form

$$\mathrm{d}\mathbf{x} = f(\mathbf{x},t)\mathrm{d}t + \mathbf{B}(\mathbf{x},t)(\mathbf{u}\mathrm{d}t + \mathrm{d}\xi) \ , \qquad (1)$$

where $\mathrm{d}\xi$ is a multivariate Wiener process with $\mathbb{E}\left[d\xi^2\right] = \mathbf{Q}(\mathbf{x},t)dt$, and $f$, $\mathbf{B}$ and $\mathbf{Q}$ may be non-linear functions. In particular note that the system is linear in the controls and both noise and controls act in the same subspace. We seek the best Markov policy, i.e., $\mathbf{u}(t) = \pi(\mathbf{x}(t),t)$, with respect to an objective of the form

$$J^\pi(\mathbf{x},t) = \mathbb{E}_{X^\pi(\cdot)|\mathbf{x}}\Big[C_\bullet(X^\pi(T)) + \int_t^T C(X^\pi(s),s) + \mathbf{u}(s)^T\mathbf{H}\mathbf{u}(s)\ \mathrm{d}s\Big] \ , \qquad (2)$$

where $T$ is some given terminal time, $C_\bullet$ and $C$ are a terminal and running cost respectively. The expectation is taken w.r.t. $X^\pi(\cdot)$, the paths of (1) starting in $\mathbf{x}$ and following policy $\pi$. The quadratic control cost, given by $\mathbf{H} \in \mathbb{R}^{d_u \times d_u}$, is further constrained by requiring it to satisfy $\mathbf{Q} = \lambda\mathbf{B}\mathbf{H}^{-1}\mathbf{B}^T$ for some constant scalar $\lambda > 0$.

It can be shown that for problems of this form, the optimised objective can be expressed as [Kappen, 2011]

$$J^*(\mathbf{x},t) = \min_\pi J^\pi(\mathbf{x},t) = -\lambda \log\Psi(\mathbf{x},t) \ , \qquad (3)$$

where $\Psi$ is a state desirability function given by the path integral

$$\Psi(\mathbf{x},t) = \mathbb{E}_{X^0(\cdot)|\mathbf{x}}\left[e^{-\int_t^T \frac{1}{\lambda}C(X^0(s),s)\mathrm{d}s}\Psi(X^0(T),T)\right] \ , \qquad (4)$$

with $\Psi(\cdot,T) = \exp\{-C_\bullet(\cdot)/\lambda\}$. The expectation in (4) is taken w.r.t. uncontrolled path of the dynamics (1), i.e. those under the policy $\pi^0(\cdot,\cdot) = 0$, starting in $\mathbf{x}_t$.

As a consequence of linear control with quadratic control cost and (3), the optimal policy $\pi^*(\mathbf{x},t)$ can be expressed directly in terms of $\Psi$ as

$$\pi^*(\mathbf{x},t) = -\mathbf{H}^{-1}\mathbf{B}(\mathbf{x})^T\nabla_\mathbf{x}J^*(\mathbf{x},t) \qquad (5)$$

$$= \mathbf{H}^{-1}\mathbf{B}(\mathbf{x})^T\frac{\lambda\nabla_\mathbf{x}\Psi(\mathbf{x},t)}{\Psi(\mathbf{x},t)} \ , \qquad (6)$$

making obtaining $\Psi$ the main computational challenge for problems in this class.

Assuming we are only interested in the optimal controls at certain time points, say $\{t_{1,\ldots,n}\}$ with $t_n = T$, it is sufficient to compute the set $\Psi_i(x) = \Psi(x,t_i)$ and (4) admits a representation in terms of the finite dimensional distribution $X = (X^0(t_0),\cdots,X^0(t_n))$. Specifically using the Markov property of $X^0(t)$ and marginalising intermediate states we obtain the recursive expression

$$\Psi_i(x_{t_i}) = \mathbb{E}_{X_{i+1}|x_{t_i}}\left[\Phi_i(x_{t_i},X_{i+1})\ \Psi_{i+1}(X_{i+1})\right] \ , \qquad (7)$$

where,

$$\Phi_i(x_{t_i},x_{t_{i+1}}) = \mathbb{E}_{X^0(\cdot)|x_{t_i},x_{t_{i+1}}}\left[e^{-\frac{1}{\lambda}\int_{t_i}^{t_{i+1}} C(X^0(s),s)\mathrm{d}s}\right] \qquad (8)$$

with the expectation taken w.r.t. uncontrolled paths from $x_{t_i}$ to $x_{t_{i+1}}$. Note that $-\lambda\log\Phi_i$ can be seen as the (optimal) expected cost for the problem of going from $x_{t_i}$ to $x_{t_{i+1}}$ over the time horizon $[t_i,t_{i+1}]$ under dynamics and running costs corresponding to those of the overall problem given in (2). Hence, the problem naturally decomposes into, on the one hand, a set of short horizon problems – or indeed a nested hierarchy of such $\Phi$ – and on the other hand, a set of recursive evaluations backwards in time.

## 3 Embedding of the Path Integral

We now demonstrate that (7) can be expressed in terms of linear operators in RKHS. While the exposition is necessarily short, [Hofmann *et al.*, 2008] provide a more through treatment of the theory of RKHS, while [Smola *et al.*, 2007], [Song *et al.*, 2009; 2011] and [Fukumizu *et al.*, 2011] provide the basic concepts on which we build.

We first concentrate on the evaluation of a single step, i.e., $\Psi_i$ given $\Psi_{i+1}$, and derive model-based analytical expressions for the evaluation of $\Psi_i$ in terms of certain operators in RKHS.

### 3.1 Model-based Analytical One Step Path Integral Embedding

We essentially show that, in the model-based case, we may write (7) as an inner product in a RKHS. We proceed by first expressing expectations in terms of inner products in a RKHS, adapting the basic expression to conditional expectations and finally considering conditional expectations of functions of both the conditional and conditioning variable.

In general, we denote by $\mathcal{H}^k$ the RKHS of functions $\mathcal{Z} \to \mathbb{R}$ associated with the positive semi-definite kernel $k(\cdot,\cdot)$. Further, let $\mathcal{P}^\mathcal{Z}$ be the set of random variables on $\mathcal{Z}$. Following [Smola *et al.*, 2007], we define the embedding operator $\mathcal{E}^k : \mathcal{P}^\mathcal{Z} \to \mathcal{H}^k$ by

$$\mathbb{E}_Z\left[h(Z)\right] = \langle h, \underbrace{\mathbb{E}_Z\left[k(Z,\cdot)\right]}_{:=\mathcal{E}^k[Z]}\rangle \quad \forall Z \in \mathcal{P}^\mathcal{Z}, h \in \mathcal{H}^k \ , \qquad (9)$$

which constitutes a direct extension of the standard embedding of individual elements $z \in \mathcal{Z}$ into $\mathcal{H}^k$, given by $\mathcal{E}^k[z] = k(z, \cdot)$ commonly encountered.

In the problem under consideration, the interest lies with the evaluation of $\Psi_i$ given in (7) and hence, in a suitable embedding of $X_{i+1}|x_i$ which would allow the required expectation to be expressed as an inner product in some RKHS. Although (9) can be directly applied – since for fixed $x_i$, $X_{i+1}|x_i$ is a simple random variable – it is convenient to consider a general conditional random variable $Z|y$ as a map $\mathcal{Y} \to \mathcal{P}^{\mathcal{Z}}$, yielding random variables over $\mathcal{Z}$ given a value $y \in \mathcal{Y}$, and define a conditional embedding $\mathcal{U}^{lk} : \mathcal{H}^l \to \mathcal{H}^k$ s.t.

$$\mathcal{E}^k[Z|y] = \mathcal{U}^{lk} \circ \mathcal{E}^l[y] \ . \tag{10}$$

Note that here $\mathcal{E}^l[y] = l(\cdot, y)$, i.e., the standard embedding operator of elements $y \in \mathcal{Y}$ used in kernel methods, and hence we may express conditional expectations as

$$\mathbb{E}_{Z|y}[h(Z)] = \langle h, \underbrace{\mathbb{E}_{Z|y}[k(Z,\cdot)]}_{:=\mathcal{E}^k[Z|y]=\mathcal{U}^{lk}[l(y,\cdot)]} \rangle \ . \tag{11}$$

Existence and a specific form for such an operator $\mathcal{U}$ satisfying (10) have been shown by [Song *et al.*, 2009] and we omit the details of its derivation.

Unlike $h$ in (11), the argument of the expectation in (7), specifically of $\Phi$, is not only a function of the random variable, i.e., $X_{i+1}$, but also of the conditioning $x_i$, and we can not apply (11) directly. We proceed by introducing an auxiliary random variable $\tilde{X}$ such that $P(\tilde{X}, X_{i+1}|\mathbf{x}_i) = P(X_{i+1}|\mathbf{x}_i)\delta_{\tilde{X}=\mathbf{x}_i}$ with $\delta$ the delta distribution, hence for all $h \in \mathcal{H}^k$

$$\begin{aligned}\mathbb{E}_{X_{i+1}|\mathbf{x}_i}[h(\mathbf{x}_i, X_{i+1})] &= \mathbb{E}_{X_{i+1},\tilde{X}|\mathbf{x}_i}\left[h(\tilde{X}, X_{i+1})\right] \\ &= \langle h, \mathcal{E}^k\left[X_{i+1}, \tilde{X}|\mathbf{x}_i\right]\rangle \ . \end{aligned} \tag{12}$$

Note that treating $\mathbf{x}_i$ as constant leads to an alternative formulation. This, although equivalent in the analytical setting, does however not immediately yield a practical empirical estimator.

We may now turn to substitution of the specific argument encountered in (7) for the generic function $h$. Specifically, assume $\mathcal{H}^\psi, \mathcal{H}^\phi$, s.t. $\Psi \in \mathcal{H}^\psi$, $\Phi \in \mathcal{H}^\phi$, are given[1]. To account for the mismatch in the arity of functions in these spaces, $\mathcal{H}^\psi$ may be trivially extended to $\mathcal{H}^{\psi'}$, a space of functions $\mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$, using the kernel $\psi'((u,v),(u',v')) = \psi(u,u')$, i.e., we identify $\mathcal{H}^\psi$ and its tensor product with the RKHS of constant functions. Hence, taking the embedding of $X_{i+1}, \tilde{X}|x_i$ into $\mathcal{H}^w = \mathcal{H}^\phi \otimes \mathcal{H}^{\psi'}$ in which the product function of $\Phi_i$, $\Psi_{i+1}$ resides, we rewrite (7), using (12) and (11), as

$$\begin{aligned}\Psi_i(\mathbf{x}) &= \mathbb{E}_{X_{i+1}|X_i=\mathbf{x}}[\Phi_i(X_{i+1},\mathbf{x}) \cdot \Psi_{i+1}(X_{i+1})] \\ &= \left\langle \Phi_i \otimes \Psi_{i+1}, \mathcal{E}^w\left[X_{i+1}, \tilde{X}|X_i = \mathbf{x}\right]\right\rangle \\ &= \left\langle \Phi_i \otimes \Psi_{i+1}, \mathcal{U}^{wk} \circ \mathcal{E}^k[\mathbf{x}]\right\rangle \ , \end{aligned} \tag{13}$$

where $k$ is some kernel over $\mathbb{R}^{d_x}$ of our choosing. As will become apparent in the following (see (16)), it is convenient for computational reasons to take $k$ to be $\psi$ as it allows for reuse of pre-computed matrices over the recursive evaluation of estimates of $\Psi$.

```
Algorithm 1: Basic Backward Algorithm
input : kernel $\psi$, transition sample $\{\mathcal{X}, \mathcal{Y}\}$, task $\Phi_{1\dots n}$
output: $\{\alpha_{0\dots n}\}$ s.t. $\hat{\Psi}_i(\mathbf{x}) = \mathbf{G}^\psi_{\mathbf{x}\mathcal{X}}\alpha_i$

$\mathbf{W} \leftarrow (\psi(\mathcal{X},\mathcal{X}) + \epsilon\mathbf{I})^{-1}$;
$\mathbf{G}_{yx} \leftarrow \psi(\mathcal{Y},\mathcal{X})$;
$\alpha_n \leftarrow \mathbf{W}\Phi_n(\mathcal{X})$;
for $i \leftarrow n-1$ to $0$ do
    $\mathbf{G}_c \leftarrow \Phi_i(\mathcal{X},\mathcal{Y})$;
    $\alpha_i \leftarrow \mathbf{W}(\mathbf{G}_c \odot \mathbf{G}_{yx}\alpha_{i+1})$;
```

## 3.2 Model-free Finite Sample Estimates

Evaluation of the embedding of random variables and in our case $\mathcal{U}$ – thus also of the path integral embedding (13) – requires evaluation of expectations of kernels and remains therefore, in most cases, intractable. Further it requires exact analytical knowledge of the system dynamics (1). However, as we shall see it is straightforward to form empirical estimates, leading to practical algorithms.

Specifically let $\mathcal{D} = \{(x,x')_{1\dots m}\}$ be a set of i.i.d. state transition samples of the uncontrolled dynamics (c.f. Section 2), e.g., a sample set obtained from trajectory executions under the policy $\pi^0$. It can be shown [Song *et al.*, 2009], that a regularized estimate of $\mathcal{U}^{kw}$ is given by

$$\hat{\mathcal{U}}^{\psi w} = \mathbf{g}^w_{\mathcal{D}}(\mathbf{G}^\psi_{\mathcal{X}\mathcal{X}} + \epsilon m\mathbf{I})^{-1}\mathbf{g}^\psi_{\mathcal{X}} \ , \tag{14}$$

where $\epsilon$ represents a regularization parameter and $\mathbf{g}^\psi_{\mathcal{X}}$,$\mathbf{g}^w_{\mathcal{X}'}$ and $\mathbf{G}^\psi_{\mathcal{X}\mathcal{X}}$ represent the vectors of embeddings and Gramian on the sample data $\mathcal{D}$ respectively, i.e., $[\mathbf{g}^\psi_{\mathcal{X}}]_i = \phi(x_i,\cdot)$ and $[\mathbf{G}^\psi_{\mathcal{X}\mathcal{X}}]_{ij} = \psi(x_i,x_j)$. In order to evaluate (13), we furthermore require the representations of $\Phi_i$ and $\Psi_{i+1}$ in their respective RKHSs. As we evaluate $\Psi$ recursively, we may assume the empirical estimate of $\Psi_{i+1}$ to be already given as $\bar{\Psi}_{i+1} = \sum_{x_j \in \mathcal{X}}[\alpha_{i+1}]_j\psi(x_j,\cdot) = \mathbf{g}^\psi_{\mathcal{X}}\alpha_i$ where $\alpha_{i+1}$ is a vector of weights. Similarly we will assume the representation of $\Phi_i$ in $\mathcal{H}^\phi$ to be given by $\mathbf{g}^\phi_{\mathcal{B}}\beta$ for some weights $\beta$ and set $\mathcal{B}$. While such a representation is, under the assumption $\Phi_i \in \mathcal{H}^\phi$, guaranteed to exists we will see that it is not necessary to explicitly compute it. Substituting the empirical operator $\hat{\mathcal{U}}^{kw}$ of (14), together with the kernel expansions of $\Phi_i$ and $\Psi_{i+1}$ into (13), matrix algebra yields the empirical estimate of $\Psi_i(\mathbf{x})$ as

$$\bar{\Psi}_i(\mathbf{x}) = \langle \psi(\mathbf{x},\cdot), \mathbf{g}^\psi_{\mathcal{X}}\alpha_i \rangle = \mathbf{G}^\psi_{\mathbf{x}\mathcal{X}}\alpha_i \tag{15}$$

with weights $\alpha_i$ given by

$$\alpha_i = [\ \underbrace{\mathbf{G}^\phi_{\mathcal{D}\mathcal{B}}\beta}_{=\Phi_i(\mathcal{X},\mathcal{X}')}\ \odot \underbrace{\mathbf{G}^\psi_{\mathcal{X}'\mathcal{A}}\alpha_{i+1}}_{=\bar{\Psi}_{i+1}(\mathcal{X}')}]^T(\mathbf{G}^\psi_{\mathcal{X}\mathcal{X}} + \epsilon m\mathbf{I})^{-1} \ , \tag{16}$$

[1] n.b., $\mathcal{H}^\phi$ is a space of functions $\mathbb{R}^{d_x} \times \mathbb{R}^{d_x} \to \mathbb{R}$, while $\mathcal{H}^\psi$ contains functions $\mathbb{R}^{d_x} \to \mathbb{R}$

where $\odot$ denotes the Hadamard product. While (16) provides the means for recursive evaluation of the weights $\alpha$, it remains to obtain corresponding representation of $\Phi_i$ given by $\beta$. However, we note the term involving the representation of $\Phi_i$ can be written as

$$\mathbf{G}^{\phi}_{\mathcal{DB}}\beta = \Phi_i(\mathcal{X},\mathcal{X}') = (\Phi_i(x_1,x_1'),\Phi_i(x_1,x_2'),\dots)^T \;. \tag{17}$$

This means obtaining an explicit representation of $\Phi$ in terms of the kernel expansion in $\mathcal{H}^{\phi}$ is not necessary, rather it is sufficient to be able to evaluate $\Phi_i$ at the sample data $\mathcal{D}$ – equivalent to evaluating the cost function (cf. (8)).

Importantly, note that $\bar{\Psi}_i$ is a finite weighted sum of kernels, hence, $\bar{\Psi}_i \in \mathcal{H}^{\psi}$, which directly allows a recursive computation of all $\bar{\Psi}_{1\dots n}$. Furthermore, all required matrices are functions of the sample data only and, as such, can be precomputed. The resulting algorithm for computation of the desirability function $\bar{\Psi}$ is summarised in Algorithm 1. Using (5), it is then straightforward to obtain an approximate optimal policy for fine discretisations of the problem.

# 4 Efficient Estimators

The basic estimator (16) has several drawbacks. For one it has a relatively high computational complexity of $\mathcal{O}(m^3)$ for the matrix inversion, only required once if the same $\mathcal{D}$ is used in each time step, and subsequently $\mathcal{O}(m^2)$ per iteration. Additionally, sample data under the uncontrolled dynamics is required, thus not allowing for off-policy learning. However, these problems can be addressed with alternative estimators for $\mathcal{U}$ based on low rank approximations [Song *et al.*, 2011] or importance sampling [Rawlik *et al.*, 2012b] We choose to ommit the discussion of these in order to address a, in our opinion, often overlooked aspect of efficiency when solving varying problems under the same dynamics. In practice, tasks are not performed in isolation, rather varying instances of often related problems have to be solved repeatedly, e.g., an optimized single reaching movement is of limited use since complex interactions require a series of such movements with changing start and target states. Previous approaches generally assume re-initialisation for each problem instance, e.g., Monte Carlo methods require novel samples, even under such trivial changes as the start state. In the following, we discuss extensions to the proposed method which improve sampling efficiency in exactly these cases, allowing efficient sample reuse over repeated applications.

## 4.1 Transfer Learning via Transition Sample Re-use

A limitation of the estimator arising in practice is the necessity of evaluating $\Phi$ at the training transitions (cf. (16) and (17)) which, in general, may be infeasible. It is therefore desirable to obtain an estimator based on evaluation of $\Phi$ on a separate, ideally arbitrary, data set $\mathcal{D}'$. Observe that

$$\begin{aligned}\mathbf{G}^{\phi}_{\mathcal{DB}}\beta &= \langle\Phi,\phi(\mathcal{D},\cdot)\rangle = \langle\Phi,\mathcal{C}^{\phi\phi}_{ZZ}\left(\mathcal{C}^{\phi\phi}_{ZZ}\right)^{-1}\phi(\mathcal{D},\cdot)\rangle \\ &\approx \underbrace{\beta^T\mathbf{G}^{\phi}_{\mathcal{BD}'}}_{\Phi(\mathcal{D}')}(\mathbf{G}^{\phi}_{\mathcal{D}'\mathcal{D}'}+\epsilon m'\mathbf{I})^{-1}\mathbf{G}^{\phi}_{\mathcal{D}'\mathcal{D}} \;,\end{aligned}$$

where $Z$ is some free random variable with support on $\mathbb{R}^{d_x}\times\mathbb{R}^{d_x}$ and we used an empirical estimator based on a data set $\mathcal{D}'=\{(x,x')_{1\dots m'}\}$ of i.i.d. samples from $Z$ (often in practice $\mathcal{D}'\subseteq\mathcal{D}$). As indicated evaluation of the r.h.s. only requires evaluation of $\Phi$ at elements of $\mathcal{D}'$, hence substituting into (16) gives the desired result. In particular we are now able to pre-compute and re-use the inverse matrix of (16) across changing tasks and, assuming time stationary dynamics, across different time steps. This is of importance for efficient estimation in, e.g., the RL setting where incurred costs are known only at observed transitions or in cases where $\Phi$ can be freely evaluated but it is expensive to do so, while generating large sets of transition samples may be comparatively cheap, e.g., the case of simple kinematic control where cost evaluation requires collision detection. Note that this form makes explicit use of the kernel $\phi$, and while we may not be able to guarantee $\Phi\in\mathcal{H}^{\phi}$, by choosing a kernel such that the projection of $\Phi$ onto $\mathcal{H}^{\phi}$ is close to $\Phi$, we can expect good results.

## 4.2 Task Augmented Sampling

We now turn to the question of the sampling distribution. While, in general, samples are required from the task agnostic dynamics $X^0$, a task often induces regularities which suggests more suitable sampling distributions. In particular, considering the role $\Phi$ takes in (16) (via (17)) as a weight vector, it appears desirable, akin to importance sampling, to concentrate samples in regions of high $\Phi$. Obviously $\Phi$ can be used to guide the choice of the sampling distribution, however, in the context of repeated tasks we can go further and incorporate $\Phi$ partly into the sampling process allowing, amongst others, for incremental learning of the task.

Consider the specific situation where one wishes to execute several task instances of a generic skill. This situation is often characterised by an invariant cost component relating to the skill and a task specific cost component – if one looks at walking as an example, we wish to stay balanced in each step but the foot placement target will differ from step to step. Formally assume the state cost decomposes as

$$C(x,\theta,t) = C_{skill}(\mathbf{x},t) + C_{task}(\mathbf{x},\theta,t) \;, \tag{18}$$

where $\theta$ parametrises the task. In this case, we may write the path integral (4) as

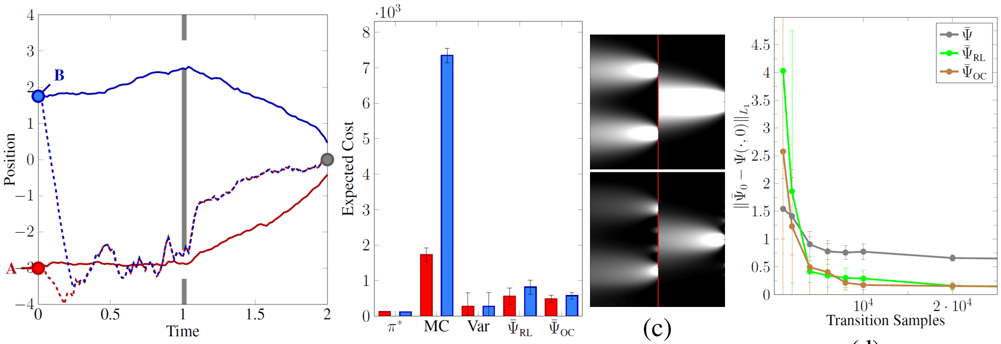
$$\Psi = \mathbb{E}_{X^{\nu}(\cdot)|x_t}\left[e^{-\int_t^T \frac{1}{\lambda}C_{task}(X^{\nu}(t),\theta,t)}\Psi(X^{\nu}(T),T)\right] \;, \tag{19}$$

where the expectation is now taken w.r.t. path of $X^{\nu}$, which are the dynamics under the optimal policy under the invariant skill component of the cost. This allows for both, incremental learning and, using the results of Section 4.1 the transfer of samples between varying tasks sharing a skill component.

# 5 Experimental Validation

## 5.1 Double Slit

We first consider the double slit problem, previously studied by [Kappen, 2005] to demonstrate Monte Carlo approaches to path integral control. The problem is sufficiently simple to allow for a closed form solution for $\Psi$ to be obtained, but



(a) (b) (c) (d)

Figure 1: Results for the double slit problem. (**a**) Problem setup and mean trajectories from policies MC and $\bar{\Psi}_{\text{RL}}$ for two start points are shown. Obstacles and target are shown in gray. (**b**) Empirical expected cost for policies based on various methods for the two start states. (**c**) The true $\Psi$ *(top)* and the estimate $\bar{\Psi}_{\text{OC}}$ *(bottom)* based on $10^4$ samples. (**d**) The $L_1$ error of estimates of $\Psi(\cdot, 0)$ as a function of (transition) sample size, n.b. in case of $\bar{\Psi}$ and $\bar{\Psi}_{\text{RL}}$ data was sampled as 100 step trajectories, for various estimators.

complex enough to highlight the shortcomings of some previous approaches. The task concerns a particle moving with constant velocity in one coordinate, while noise and controls affects it's position in an orthogonal direction. The aim is to minimise the square error to a target position at some final time, while also avoiding obstacles at some intermediate time, as illustrated in Fig. 1(a). Specifically, the one dimensional dynamics are $\mathrm{d}x = u + \mathrm{d}\xi$ and the cost is given by

$$C_{\bullet}(x) = \omega(x - x_{target})^2$$
$$C(x,t) = \begin{cases} 10^4 & \text{if } t = \frac{T}{2} \text{ and } x \in \mathit{Obstacle} \\ 0 & \text{else} \end{cases} ,$$

where $\omega$ is a weight. We considered a discretisation with time step $0.02s$, i.e. 100 time steps.

We compare the true optimal policy to those obtained using two variants of the proposed estimator, $\bar{\Psi}_{\text{OC}}$ and $\bar{\Psi}_{\text{RL}}$. The latter is based on a RL setting, learning from trajectory data without access to the cost, and uses the approach for sample sharing across time steps discussed in Section 4.1. Meanwhile, $\bar{\Psi}_{\text{OC}}$ is based on single transitions from uniformly sampled start states and uses knowledge of the cost function to evaluate $\Phi$ in each step. In both cases we use the low rank approximation and square exponential kernels $\psi(x, y) = \exp\{(x-y)^2/\gamma\}$ with $\gamma$ set to the median distance of the data. For comparison, we also consider two alternative approaches – firstly, the trajectory based Monte Carlo approach of [Theodorou *et al.*, 2009], using the same number of trajectories as used in the RL setting and on the other hand, a variational approximation, specifically a Laplace approximation to the true $\Psi$ to obtain a linear approximation of the optimal policy. As can be seen in Fig. 1(b), the proposed approach leads to policies which significantly improve upon those based on the alternative Monte Carlo approach and which are comparable to those obtained from the variational approximation, which however was computed based on knowledge of the true $\Psi$. In particular, note that the proposed approach makes better use of the sample provided, finding a policy which is applicable for varying starting positions, as illustrated in Fig. 1(a). As seen from the trajectories in Fig. 1(a), the Monte Carlo approach on the other hand fails to capture the multi modality of the optimal policy leading to severely impoverished results when applied to starting point B without sampling a new data set (c.f. Fig. 1(b)). The variational approximation on the other hand similarly requires recomputation for each new starting location, without which results would also be significantly affected.

To illustrate the dependence of the estimate on the sample size we compare in Fig. 1(c) the evolution of the $L_1$ error of the estimates of $\Psi$ at time $t = 0$. Sample size refers to total number of transition samples seen, hence for $\bar{\Psi}_{\text{RL}}$ the number of trajectories was the sample size divided by 100. In order to also highlight the advantages of the sample re-use afforded by the approach in Fig. 4.1, we also compare with $\bar{\Psi}$, the basic estimator given data of the same form as $\bar{\Psi}_{\text{RL}}$, i.e. recursive application of (16) without sample sharing across time steps.

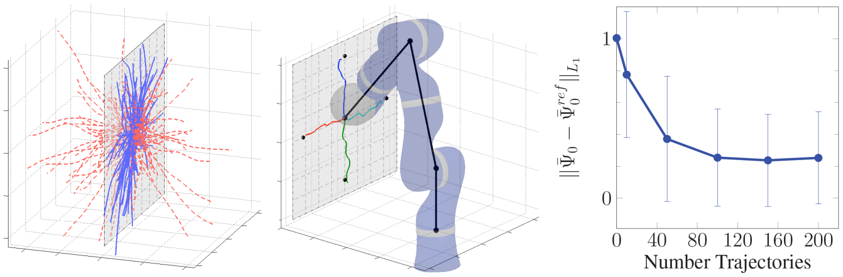
## 5.2 Arm Subspace Reaching Task

We consider reaching tasks on a subspace of the end-effector space of a torque controlled 5dof arm, simulating constrained tasks such as, for e.g., drawing on a whiteboard or pushing objects around on a table. Here the skill component consists of moving with the end-effector staying close to a two dimensional task space, while the task instances are given by specific reach targets. The task space used is a linear subspace of the end effector space, n.b., hence, a non linear subspace of the joint space, and the cost comprises the two components

$$C_{skill}(\mathbf{x}, t) = \omega_{skill}\|\mathbf{J}\varphi(\mathbf{x}) - \mathbf{j}\|^2$$
$$C_{task}(\mathbf{x}, \theta) = \omega_{task}\|\varphi(\mathbf{x}) - \theta\|^2 ,$$

where $\varphi(\cdot)$ is the mapping from joint to end-effector coordinates, $\mathbf{J}$ & $\mathbf{j}$ define the task subspace, $\theta$ specifies the reaching target and $\omega$'s are weights. We again consider position control over a 2s horizon with a 0.02s discretisation.

This task is challenging for sample based approaches as the low cost trajectories are restricted to a small subspace, necessitating large sample sizes to obtain good results for an individual reaching target, even if, as suggested by [Theodorou



(a) (b) (c)

Figure 2: Results in the reaching task. (**a**) Training trajectories under the skill augumented policy (solid blue) and $\pi^0$ (dashed red) with task space. (**b**) Illustration of the task setup and example trajectories of policies after 100 training trajectories for a set of reaching tasks. The black dots show individual reaching targets with the arm shown in it's initial pose. (**c**) The $L_1$ error of estimates of $\Psi(\cdot, 0)$ as a function of training trajectories measured with respect to an estimate trained on 5000 trajectories. The data point coresponding to #traj = 0 is based on the estimate is of $\bar{\Psi}$ taking only $C_{skill}$ into account (see text for details).

*et al.*, 2009] and done here, an inverse dynamics policy is used which significantly improves end-effector exploration. However, concentrating on the case of changing targets, we exploit the ideas from Section 4.2 by assuming the operators have been estimated under the skill augmented dynamics[2] (cf. (19)), and consider subsequent learning for a novel task using the estimator from Section 4.1, utilising the already estimated operators in two ways. On the one hand, they are directly used in the calculation of $\bar{\Psi}$, on the other hand, noting, that as the trajectories are only required to provide $\mathcal{D}'$, hence do not have to be sampled under a specific policy, we use the policy arising when considering $C_{skill}$ only, i.e., the skill policy associated with $\bar{\Psi}$ computed using the given operators and $C_{task}(\cdot) = 0$.

The advantage of sampling under the skill policy is illustrated in Fig. 2(a) where sample trajectories under both the skill and null policy are shown, demonstrating that the former more effectively explores the task relevant sub space. Mean trajectories for policies learned from 100 trajectories for a set of tasks are illustrated in Fig. 2(b). In Fig. 2(c) we plot the $L_1$ error of $\bar{\Psi}$ as a function of trajectories averaged over ten $\theta$. As the true $\Psi$ is not available for this task we show the error w.r.t. a $\bar{\Psi}_0^{ref}$ computed from 5000 trajectories, principally to illustrate the rapid convergence of the estimator.

## 6 Conclusion

We have presented a novel approach for solving stochastic optimal control problems which are of the path integral control form using Monte Carlo estimates of operators arising from a RKHS embedding of the problem, leading to a consistent estimate of $\Psi$. While direct application of Monte Carlo estimation to point evaluation of $\Psi$ also yields a consistent estimate, it is impractical for computation of controls for anything but simple problems, requiring a trajectory sample for each state at which an action is to be computed. Although previous work, e.g. by [Theodorou *et al.*, 2009; 2010], has suggested approaches to overcome the problem of sample complexity, these sacrifice consistency in the process and we demonstrate that the proposed approach significantly improves upon them in terms of generalization to a policy (cf. results in Fig. 1(a,b)). We furthermore show that the presented estimators allow for sample re-use in situations which previously required an entirely novel sample set. In particular we consider transfer in cases where execution of several, potentially related, tasks on the same plant is required, demonstrating that it is possible to exploit samples from all tasks to learn invariant aspects.

Note that as $\Phi$ itself defines a local optimal control problem, an alternative perspective on the proposed method is as a principled approach to combining solutions to local control problem to solve a more complex large scale problem. In future work we aim to elaborate on this interpretation by combining the methods presented with alternative approaches, e.g., variational methods, which may be sufficient to provide good estimates for the comparatively simpler local problems.

The choice of kernel has been largely ignored here, but one may expect improved results by making informed kernel choices based on prior knowledge about the structure of the problem.

In related work, [Grünewälder *et al.*, 2012] recently proposed to utilise RKHS embeddings of the transition probability for computation value functions in MDPs. While the methodologies are in principle comparable, our work goes beyond that of Grünewälder *et al.* in several aspects. For one, we directly obtain the optimal controls, rather then just computing the value function, leaving obtaining optimal controls to be performed by explicit maximisation. Furthermore Grünewälder *et al.* concentrate on finite state problems (where computation of the optimal u is simpler). We meanwhile, concentrate on the harder continuous problem and in particular, provide convergence guarantees in this setting. Finally, we make a contribution to the efficient estimation of the required quantities by exploiting the structure of the problem, allowing efficient sample re-use and transfer.

[2] n.b., while here such a sample is generated explicitly, the more time consuming approach of using the importance sample based estimator and collecting a sample under $X^0$ could be used

## References

[Broek *et al.*, 2010] J.L. van den Broek, W.A.J.J. Wiegerinck, and H.J. Kappen. Risk sensitive path integral control. In *UAI*, 2010.

[Broek *et al.*, 2011] J.L. van den Broek, Wiegerinck W.A.J.J., and Kappen H.J. Stochastic optimal control of state constrained systems. *International Journal of Control*, 84(3):597–615, 2011. Published.

[Friston *et al.*, 2010] K. J. Friston, J. Daunizeau, J. Kilner, and S.J. Kiebel. Action and behavior: a free-energy formulation. *Biol Cybern.*, 102(3):227–260, 2010.

[Fukumizu *et al.*, 2011] K. Fukumizu, L. Song, and A. Gretton. Kernel Bayes' rule. In *NIPS*, 2011.

[Grünewälder *et al.*, 2012] S. Grünewälder, G. Lever, L. Baldassarre, M. Pontil, and A. Gretton. Modelling transition dynamics in MDPs with RKHS embeddings. In *ICML*, 2012.

[Hofmann *et al.*, 2008] T. Hofmann, B. Schlkopf, and A.J. Smola. Kernel methods in machine learning. *The Annals of Statistics*, 36(3):1171–1220, 2008.

[Kappen, 2005] H.J. Kappen. A linear theory for control of non-linear stochastic systems. *Physical Review Letters*, 95:200201, 2005.

[Kappen, 2011] H.J. Kappen. Optimal control theory and the linear Bellman equation. In *Inference and Learning in Dynamic Models*. 2011.

[Mensink *et al.*, 2010] T. Mensink, J. Verbeek, and H.J. Kappen. EP for efficient stochastic control with obstacles. In *European Conference on Artificial Intelligence*, 2010.

[Rawlik *et al.*, 2012a] K. Rawlik, M. Toussaint, and S. Vijayakumar. On stochastic optimal control and reinforcement learning by approximate inference. In *R:SS*, 2012.

[Rawlik *et al.*, 2012b] K. Rawlik, M. Toussaint, and S. Vijayakumar. Path integral control by reproducing kernel hilbert space embedding. Technical report, arXiv:1208.2523, 2012.

[Smola *et al.*, 2007] A. Smola, A. Gretton, L. Song, and B. Schlkopf. A hilbert space embedding for distributions. In *Algorithmic Learning Theory: 18th International Conference*, 2007.

[Song *et al.*, 2009] L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions. In *ICML*, 2009.

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

[Theodorou *et al.*, 2009] E.A. Theodorou, J. Buchli, and S. Schaal. Path integral-based stochastic optimal control for rigid body dynamics. In *Adaptive Dynamic Programming and Reinforcement Learning*, 2009.

[Theodorou *et al.*, 2010] E. Theodorou, J. Buchli, and S. Schaal. A generalized path integral control approach to reinforcement learning. *Journal of Machine Learning Research*, (11):3137–3181, 2010.

[Todorov, 2007] E. Todorov. Linearly-solvable markov decision problems. In *NIPS*. 2007.

[Todorov, 2009] E. Todorov. Efficient computation of optimal actions. *PNAS*, 106:11478–11483, 2009.

[Toussaint, 2009] M. Toussaint. Robot trajectory optimization using approximate inference. In *ICML*, 2009.

[Zhong and Todorov, 2011a] M. Zhong and E. Todorov. Aggregation methods for linearly-solvable MDPs. In *World Congress of the International Federation of Automatic Control*, 2011.

[Zhong and Todorov, 2011b] M. Zhong and E. Todorov. Moving least-squares approximations for linearly-solvable stochastic optimal control problems. *Journal of Control Theory and Applications*, 9:451–463, 2011.