Path Integral Control by Reproducing Kernel Hilbert Space Embedding

Konrad Rawlik
School of Informatics
University of Edinburgh
Edinburgh, UK

Marc Toussaint
FU Berlin
Berlin, Germany

Sethu Vijayakumar
School of Informatics
University of Edinburgh
Edinburgh, UK

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 and Reinforcement Learning problems remains challenging, some recent work [7] 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, e.g., [15, 1]. 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 it's 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 [15, 16]. Additionally, the approach remains model free, allowing it's application to the Reinforcement Learning setting. This is in contrast to variational [9] or function approximation [20, 21] 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 [17], inference control [19] and free energy control [3] all require solving an underlying problem of equivalent form, making the methods proposed directly applicable in these contexts. Furthermore [10] 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 [18], as well as risk sensitive control [2] also exist.

2   Path Integral Control

In this section we briefly review the path integral approach to stochastic optimal control [7], for a more detailed treatment, see [8, 15]. 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

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

where $d\xi$ 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 affine 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

$$J^\pi(\mathbf{x}, t) = \mathbb{E}_{X^\pi(t \to T)|\mathbf{x}} \left[C_\bullet(X^\pi(T)) + \int_t^T C(X^\pi(s), s) + \mathbf{u}(s)^T \mathbf{H}\mathbf{u}(s)ds \right] , \tag{2}$$

where T is some given terminal time and the expectation is taken w.r.t. to path of (1) starting in $\mathbf{x}$ and following policy π. The control cost 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

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

where Ψ is given by the path integral

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

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 Ψ as

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

making obtaining Ψ the main computational challenge for problems in this class.

Assuming we are only interested in the optimal controls at certain time points, say $\{t_{1,\dots,n}\}$ 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

$$\Psi_i(x_{t_i}) = \mathbb{E}_{X_{i+1}|x_{t_i}} \left[\Phi_i(x_{t_i}, X_{i+1}) \cdot \Psi_{i+1}(X_{i+1}) \right] . \tag{6}$$

Here,

$$\Phi_i(x_{t_i}, x_{t_{i+1}}) = \mathbb{E}_{X^0(t_i \to t_{i+1})|x_{t_i}, x_{t_{i+1}}} \left[e^{-\frac{1}{\lambda} \int_{t_i}^{t_{i+1}} C(X^0(s), s)ds} \right] , \tag{7}$$

where the expectation is taken w.r.t. uncontrolled path 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 Φ – and on the other hand, a set of recursive evaluations backwards in time.

3 Embedding of the Path Integral

We now demonstrate that (6) can be expressed in terms of linear operators in RKHSs. While the exposition is necessarily short, [6] provides a more through treatment of the theory of RKHSs while [11, 14, 12, 4] provide the basic concepts on which we build.

3.1 Analytical One Step Path Integral Embedding

Let $\mathcal{H}^k$ denote the reproducing kernel Hilbert space 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 [11], we define the embedding operator $\mathcal{E}^k : \mathcal{P}^{\mathcal{Z}} \to \mathcal{H}^k$ by

$$\langle h, \mathcal{E}^k [Z] \rangle = \mathbb{E}_Z [h(Z)] \quad \forall Z \in \mathcal{P}^{\mathcal{Z}}, h \in \mathcal{H}^k , \tag{8}$$

which constitutes a direct extension of the standard embedding of individual elements $z \in \mathcal{Z}$ into $\mathcal{H}^k$ used more commonly in the literature.

In the problem under consideration, the interest lies with the evaluation of Ψ_i given in (6) 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 (8) 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 operator $\mathcal{U}^{lk} : \mathcal{H}^l \to \mathcal{H}^k$ s.t.

$$\mathcal{E}^k [Z|y] = \mathcal{U}^{lk} \circ \mathcal{E}^l [y] , \tag{9}$$

where $\mathcal{E}^l [y] = l(\cdot, y)$, i.e., the standard embedding operator of elements $y \in \mathcal{Y}$ used in kernel methods. An explicit form of the operator $\mathcal{U}$ is given in [14] by means of *covariance operators*, which are generalizations of covariance matrices. Specifically, the uncentered covariance operator $\mathcal{C}_{ZY}^{kl}$ for the joint random variable (Z, Y) is given by

$$\mathcal{C}_{ZY}^{kl} = \mathbb{E}_{(Z,Y)} [k(Z, \cdot) \otimes l(Y, \cdot)] , \tag{10}$$

where $\otimes$ denotes the tensor product. Note that we can see $\mathcal{C}_{ZY}^{kl}$ as an embedding of (Z, Y) into the tensor product space $\mathcal{H}^w = \mathcal{H}^k \otimes \mathcal{H}^l$, which is the RKHS of the product kernel $w((z, y), (z', y')) = k(z, z')l(y, y')$. Now, under certain technical considerations detailed in [4] but beyond the scope of this paper,

$$\mathcal{U}^{lk} = \mathcal{C}_{ZY}^{kl} \left(\mathcal{C}_{YY}^{ll} \right)^{-1} \tag{11}$$

satisfies (9).

However, as the argument of the expectation, specifically of Φ, is not only a function of the random variable, i.e., X_{i+1}, but also of the conditioning x_i, we can not apply (9) 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 δ the delta distribution, hence

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

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 as further discussed in the supplementary material.

Now, 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 Φ_i, Ψ_{i+1} resides, using (6) and further applying (8) and (9) we have

$$\Psi_i(\mathbf{x}) = \mathbb{E}_{X_{i+1}|X_i=\mathbf{x}} [\Phi_i(\mathbf{x}, X_{i+1}) \cdot \Psi_{i+1}(X_{i+1})] \tag{13}$$

$$= \left\langle \Phi_i \otimes \Psi_{i+1}, \mathcal{E}^w \left[X_{i+1}, \tilde{X}|X_i = \mathbf{x} \right] \right\rangle \tag{14}$$

$$= \left\langle \Phi_i \otimes \Psi_{i+1}, \mathcal{U}^{wk} \circ \mathcal{E}^k [\mathbf{x}] \right\rangle , \tag{15}$$

---

[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 k is some kernel over $\mathbb{R}^{d_x}$ of our choosing. As will become apparent in the following (see (18)), it is convenient for computational reasons to take k to be ψ as it allows for re-use of pre-computed matrices over the recursive evaluation of estimates of Ψ.

3.2   Finite Sample Estimates

Evaluation of $\mathcal{U}$ – thus also of the path integral embedding (15) – requires evaluation of expectations of kernels and remains therefore, in most cases, intractable. However as the operators are expressed in terms of expectations, it is straightforward to form empirical estimates, leading to practical algorithms.

First consider the general case, given a set $\mathcal{D} = \{(z, y)_{0\ldots m}\}$ of i.i.d. samples from (Z, Y). An estimate of (10) is given by

$$\hat{\mathcal{C}}_{\mathcal{D}}^{kl} = \frac{1}{m} \sum_{i=1}^{m} k(\cdot, z_i) \otimes l(\cdot, y_i) \ . \tag{16}$$

Using the latter in conjunction with (11), a regularized estimate of $\mathcal{U}^{lk}$ is given by

$$\hat{\mathcal{U}}_{\mathcal{D}}^{lk} = \mathbf{g}_{\mathcal{Z}}^{k}(\mathbf{G}_{\mathcal{Y}\mathcal{Y}}^{l} + \epsilon m \mathbf{I})^{-1}\mathbf{g}_{\mathcal{Y}}^{l} \ , \tag{17}$$

where ϵ represents a regularization parameter and $\mathbf{g}_{\mathcal{A}}^{k}$, $\mathbf{G}_{\mathcal{A}\mathcal{B}}^{k}$ represents the vector of embeddings and Gramian respectively, i.e. $[\mathbf{g}_{\mathcal{A}}^{k}]_i = k(a_i, \cdot)$ and $[\mathbf{G}_{(\mathcal{A},\mathcal{B})}^{k}]_{ij} = k(a_i, b_j)$, for given sets $\mathcal{A}, \mathcal{B}$ and kernel k.

Now, turning to the specific expression of interest, i.e., Ψ in (15), we can form an empirical estimate based on $\mathcal{D} = \{(x, x')_{1\ldots m}\}$ sampled i.i.d. from a joint distribution $P(X', X) = p_{\pi^0}(X'|X)\mu(X)$, s.t. $p_{\pi^0}(X'|X)$ is the p.d.f. of $X_{i+1}|X_i$ and μ is a free prior. Specifically, assume the representation of Φ_i in $\mathcal{H}^{\phi}$ is $\mathbf{g}_{\mathcal{B}}^{\phi}\beta$, which we do not assume to be finite dimensional. Then, given a empirical estimate $\bar{\Psi}_{i+1} = \mathbf{g}_{\mathcal{A}}^{\psi}\alpha_{i+1}$, based on some set $\mathcal{A}$, we obtain the estimate

$$\bar{\Psi}_i = \mathbf{g}_{\mathcal{X}}^{\psi}\alpha_i \qquad \text{with} \qquad \alpha_i = \left[\mathbf{G}_{\mathcal{D}\mathcal{B}}^{\phi}\beta \odot \mathbf{G}_{\mathcal{X}'\mathcal{A}}^{\psi}\alpha_{i+1}\right]^{T} (\mathbf{G}_{\mathcal{X}\mathcal{X}}^{\psi} + \epsilon m \mathbf{I})^{-1} \ , \tag{18}$$

where $\odot$ denotes the Hadamard product. The term $\mathbf{G}_{\mathcal{D}\mathcal{B}}^{\phi}\beta$ takes– assuming without loss of generality, $\Phi \in \mathcal{H}^{\phi}$– the particularly simple form

$$\mathbf{G}_{\mathcal{D}\mathcal{B}}^{\phi}\beta = \Phi(\mathcal{X}, \mathcal{X}') = (\Phi(x_1, x_1'), \Phi(x_1, x_2'), \ldots)^{T} \ . \tag{19}$$

Hence, obtaining an explicit representation of Φ, or indeed choosing $\mathcal{H}^{\phi}$, is not necessary.

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 $\bar{\Psi}_1 \ldots n$ and leads, using (5), to an approximate optimal policy for fine discretisations of the problem. Furthermore, all required matrices are functions of the sample data only and as such can be pre-computed. Finally, the estimator is consistent (see supplementary material for proof). While for bounded expected costs, convergence of $\bar{\Psi}$ implies convergence of the estimate of the expected cost (see supplementary material), convergence of the latter can be slow for large values due to the log transform, leading in practice to poor policies in regions where Ψ is small. We would like to emphasize that this problem is not limited to the methods proposed here, but is a characteristic of any approach based on estimation of Ψ, e.g., as also noted by [21]. To overcome this problem in practice, we form a Laplace approximation to $\bar{\Psi}$ at a local mode and use it where $\bar{\Psi}$ is small - this corresponds to a local quadratic approximation of the value function, resulting in a linear policy which steers the system towards regions of high $\bar{\Psi}$.

4   Efficient Estimators

The basic estimator (18) 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

4

the uncontrolled dynamics is required, thus not allowing for off-policy learning. To overcome these problems two alternative estimators based on weighted samples, which partly address these issues, are discussed in the supplementary material. Specifically, the estimator employs Gram-Schmidt orthogonalisation, presented previously by [12], which reduces the computational complexity to $\mathcal{O}(\hat{m}^2)$ with $\mathcal{O}(\hat{m}^3 + \hat{m}^2 m)$ pre-computations for a chosen $\hat{m} \ll m$, and a novel importance sampling based estimator. We choose to defer 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 re-use 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 Φ at the training transitions (cf. (18) and (19)) which, in general, may be infeasible. It is therefore desirable to obtain an estimator based on evaluation of Φ on a separate, ideally arbitrary, data set $\mathcal{D}'$. Observe that

$$\mathbf{G}^\phi_{\mathcal{D}\mathcal{B}}\beta = \langle \Phi, \phi(\mathcal{D}, \cdot) \rangle = \langle \Phi, \mathcal{C}^{\phi\phi}_{ZZ} \left(\mathcal{C}^{\phi\phi}_{ZZ} \right)^{-1} \phi(\mathcal{D}, \cdot) \rangle \approx \overbrace{\beta^T \mathbf{G}^\phi_{\mathcal{B}\mathcal{D}'}}^{\Phi(\mathcal{D}')} (\mathbf{G}^\phi_{\mathcal{D}'\mathcal{D}'} + \epsilon m' \mathbf{I})^{-1} \mathbf{G}^\phi_{\mathcal{D}'\mathcal{D}} \ ,$$

where Z is an 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 $\mathcal{D}' = \{(x, x')_{1\ldots m'}\}$ 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 Φ at elements of $\mathcal{D}'$, hence substituting into (18) gives the desired result. In particular we are now able to pre-compute and re-use the inverse matrix of (18) across changing tasks and, assuming time stationary dynamics, across different time steps. This is of importance for efficient estimation in, e.g., the Reinforcement Learning setting where incurred costs are known only at observed transitions or in cases where Φ 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 ϕ, and while we may not be able to guarantee $\Phi \in \mathcal{H}^\phi$, by choosing a kernel such that the projection of Φ onto $\mathcal{H}^\phi$ is close to Φ, 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 Φ takes in (18) as a weight vector, it appears desirable, akin to importance sampling, to concentrate samples in regions of high Φ. Obviously Φ can be used to guide the choice of the prior μ (c.f. Section 3.2), however, in the context of repeated tasks we can go further and incorporate Φ 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

$$C(x, \theta, t) = C_{skill}(\mathbf{x}, t) + C_{task}(\mathbf{x}, \theta, t) \ , \tag{20}$$

where θ parameterises the task. In this case, we may write the path integral (4) as

$$\Psi = \mathbb{E}_{X^\nu(t \to T)|x_t} \left[e^{-\int_t^T \frac{1}{\lambda} C_{task}(X^\nu(t), \theta, t)} \Psi(X^\nu(T), T) \right] \ , \tag{21}$$

5

where the expectation is now taken w.r.t. path of X^ν, which are modified dynamics which absorb the invariant skill component of the cost, i.e., they bias the path dynamics based on the Fokker-Plank equation

$$\partial_t \nu = -\frac{C_{skill}}{\lambda}\nu - \nabla_x(f\nu) + \nabla_x^2(\mathbf{B}\mathbf{B}^T\nu) \ , \tag{22}$$

in other words, the augmented dynamics tends to restrict the solutions to lie on, or at least stay close to, some *skill space.*

A practical approach for exploiting the induced structure, is to learn the relevant subspace from a few example demonstrations sampled, using e.g. the approach in [5], and sample $\mathcal{D}$ on the learned space. Such explicit learning of the space has several advantages; foremost, we can use knowledge of the space to choose an appropriate kernel. Also, while C_{task} is generally well defined by specific objectives we wish to achieve, the skill component often takes a more abstract form, e.g. we may desire movements to overall appear 'natural', and may only be given implicitly by expert demonstrations of desired movements, in which case the proposed framework allows (21) to be used to perform optimal control without explicitly referring to the implicit costs.

5   Experimental Validation

5.1   Double Slit

We first consider the double slit problem, previously studied by [7] to demonstrate Monte Carlo approaches to path integral control. The problem is sufficiently simple to allow for a closed form solution for Ψ to be obtained, but 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 $dx = u + d\xi$ and the cost is given by

$$C_\bullet(x) = \omega(x - x_{target})^2 \quad \text{and} \quad C(x,t) = \begin{cases} 10^4 & \text{if } t = \frac{T}{2} \text{ and } x \in Obstacle \\ 0 & \text{else} \end{cases} , \tag{23}$$

where ω 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, $\bar{\Psi}_{OC}$ and $\bar{\Psi}_{RL}$. The latter is based on a Reinforcement learning 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, $\bar{\Psi}_{OC}$ is based on single transitions from uniformly sampled start states and uses knowledge of the cost function to evaluate Φ in each step. In both cases we use the low rank approximation (see supplementary material) and square exponential kernels $\psi(x,y) = \exp\{(x-y)^2/\lambda\}$ with λ set to the median distance of the data. For comparison, we also consider two alternative approaches – firstly, the trajectory based Monte Carlo approach of [16], using the same number of trajectories as used in the Reinforcement Learning setting and on the other hand, a variational approximation, specifically a Laplace approximation to the true Ψ 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 Ψ. 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 (cf. Fig. 1(b)). The variational approximation on the other hand similarly requires re-computation for each new starting location, without which results would also be significantly affected.
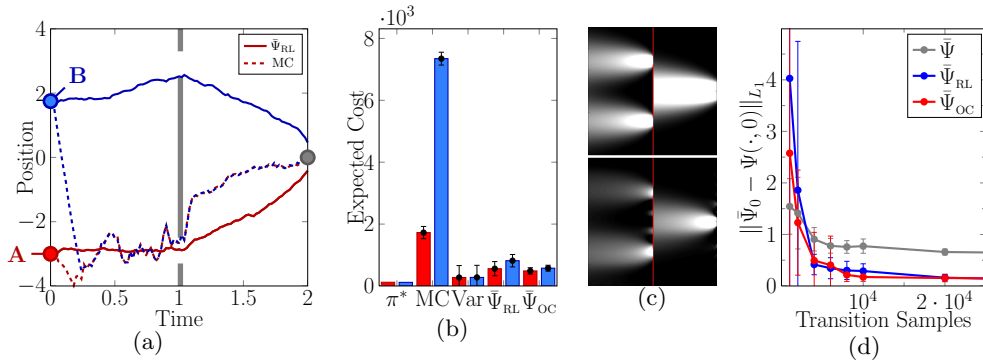
Figure 1: Results for the double slit problem. (a) Problem setup and mean trajectories from policies MC and $\bar{\Psi}_{\mathrm{RL}}$ for two start points are shown. Obstacles and target are shown in gray. (b) Empirical expected cost for policies based on various methods for the two start states. (c) The true Ψ *(top)* and the estimate $\bar{\Psi}_{\mathrm{OC}}$ *(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 $\bar{\Psi}_{\mathrm{RL}}$ data was sampled as 100 step trajectories, for various estimators.

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 Ψ at time $t = 0$. Sample size refers to total number of transition samples seen, hence for $\bar{\Psi}_{\mathrm{RL}}$ 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 Section 4.1, we also compare with $\bar{\Psi}$, the basic estimator given data of the same form as $\bar{\Psi}_{\mathrm{RL}}$, i.e. recursive application of (18) 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

$$C_{skill}(\mathbf{x}, t) = \omega_{skill} \|\mathbf{J}\varphi(\mathbf{x}) - \mathbf{j}\|^2 \quad \text{and} \quad C_{task}(\mathbf{x}, \theta) = \omega_{task}\|\varphi(\mathbf{x}) - \theta\|^2 , \qquad (24)$$

where $\varphi(\cdot)$ is the mapping from joint to end-effector coordinates, $\mathbf{J}$ & $\mathbf{j}$ define the task subspace, θ specifies the reaching target and ω'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 in [16] 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. (21)), 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 $C_{task}(\cdot) = 0$.

---

[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
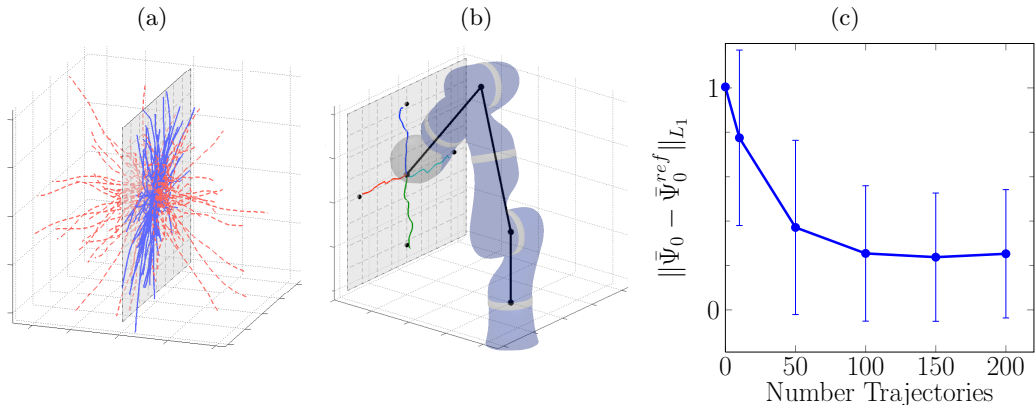
Figure 2: Results in the reaching task. (a) Training trajectories under the skill augumented policy (solid blue) and π^0 (dashed red) with task space. (b) Illustration of the task setup and example trajectories of policies after 100 training trajectories for a set of reaching tasks. The black dots show 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).

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 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 θ. As the true Ψ is not available for this task we show the error w.r.t. a $\bar{\Psi}$ 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 Ψ. While direct application of Monte Carlo estimation to point evaluation of Ψ 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., [16] and similarly [15], 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 Φ 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 persented 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.

8

References

[1] 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.

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

[3] 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.

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

[5] I. Havoutis and S. Ramamoorthy. Geodesic trajectory generation on learnt skill manifolds. In *ICRA*, 2010.

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

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

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

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

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

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

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

[13] L. Song, A. Gretton, and C. Guestrin. Nonparametric tree graphical models via kernel embeddings. In *AISTATS*, 2010.

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

[15] 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.

[16] 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.

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

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

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

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

[21] 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.

A Alternative embedding

As indicated in the main text (cf. section 3.1) an alternative representation to the embedding (15) exists. Observe that for the purposes of the expectation the conditioning variable is fixed and Φ is in fact only a function of it's second argument, making it possible to apply (9) embedding $X_{i+1}|x_i$ into the tensor space in which the product of Ψ and the partially evaluated Φ resides. Formally define the operator for partial evaluation on $\mathcal{H}^\phi$

$$\mathcal{R}_x[h] = h(x, \cdot) \quad \forall x \in \mathbb{R}^{D_\mathbf{x}}, h \in \mathcal{H}^\phi \tag{25}$$

In particular note that for $\mathcal{R}_x : \mathcal{H}^\phi \to \mathcal{H}^{\phi_\mathbf{x}}$ where $\phi_\mathbf{x} = \phi((x, \cdot), (x, \cdot))$. We can now write $\Phi_i(\mathbf{x}, \cdot) = \mathcal{R}_\mathbf{x}[\Phi_i] \in \mathcal{H}^{\phi_\mathbf{x}}$ and application of (8) and (9) to (6) leads to

$$\Psi_i(\mathbf{x}) = \mathbb{E}_{X_{i+1}|X_i=\mathbf{x}}[\mathcal{R}_\mathbf{x}[\Phi_i](X_{i+1}) \cdot \Psi_{i+1}(X_{i+1})] \tag{26}$$

$$= \langle \mathcal{R}_\mathbf{x}[\Phi_i] \otimes \Psi_{i+1}, \mathcal{E}^w[X_{i+1}|X_i = \mathbf{x}] \rangle \tag{27}$$

$$= \langle \mathcal{R}_\mathbf{x}[\Phi_i] \otimes \Psi_{i+1}, \mathcal{U}^{wk} \circ \mathcal{E}^k[\mathbf{x}] \rangle \tag{28}$$

where $\mathcal{H}^w = \mathcal{H}^{\phi_\mathbf{x}} \otimes \mathcal{H}^\psi$ and k is some kernel of our choosing on $\mathbb{R}^{D_x}$ which again we take to be ψ.

Although (28) is formally equivalent to the embedding derived in the main text, i.e., (15), the yield different empirical estimates. Specifically, applying (17) to the (28) we obtain $\hat{\Psi}_i(\mathbf{x}) = \mathbf{G}^\psi_{\mathbf{x}\mathcal{X}}\alpha(\mathbf{x})$ with

$$\alpha(\mathbf{x}) = \left[\mathbf{G}^\phi_{(\mathbf{x}\mathcal{Y})\mathcal{R}}\beta \odot \mathbf{G}^{\psi'}_{\mathcal{Y}\mathcal{X}'}\alpha'\right]^T (\mathbf{G}^\psi_{\mathcal{X}\mathcal{X}} + \epsilon n\mathbf{I})^{-1} \tag{29}$$

Hence, although this approach allows us to evaluate $\hat{\Psi}_i$ at specific points, we do not directly obtain a a finite dimensional representation of $\hat{\Psi}_i$ in some RKHS. Furthermore, due to the dependence on the evaluation point, the Gram matrix $\mathbf{G}^\phi_{(\mathbf{x}\mathcal{X})\mathcal{R}}$ can in general not be pre computed. None the less this form may have it's applications for a forward, backwards algorithm where $\mathbf{G}^\phi_{(\mathbf{x}\mathcal{X})\mathcal{R}}$ is used for selection of an active set $\mathcal{X}$ for which α's are computed in a backwards pass.

B Alternative Estimators

We now discuss the two alternative estimators based on weighted samples alluded to in the main text (cf. section 4).

B.1 Low rank Approximation

First, we address the computational complexity of (18), which is $\mathcal{O}(m^3)$ for the matrix inversion, which may be precomputed, and $O(m^2)$ for subsequent computations. Although such costs are acceptable for reasonably sized problems, they may prove prohibitive for application to realistic robotic systems. However we can apply a Gram-Schmidt orthogonalisation of $\mathbf{g}^k_\mathcal{X}$, $\mathbf{g}^k_{\mathcal{X}'}$, as proposed by [12]. Summarising we approximate $\mathbf{g}^k_\mathcal{X} \approx \mathbf{g}^k_\mathcal{Y}\mathbf{W}_x$ and $\mathbf{g}^k_{\mathcal{X}'} \approx \mathbf{g}^k_{\hat{\mathcal{X}}'}\mathbf{W}_{x'}$, where $\mathcal{Y} \subseteq \mathcal{X}$, $\mathcal{Y}' \subseteq \mathcal{X}'$ and $\mathbf{W}_x, \mathbf{W}_{x'}$ are weight matrices. Substituting into (18) we may then obtain the alternative estimator

$$\alpha_i = \left[\mathbf{G}^\phi_{\mathcal{D}'\mathcal{B}}\beta \odot \mathbf{G}^\psi_{\mathcal{Y}'\mathcal{A}}\alpha_{i+1}\right]^T \mathbf{W}_{x'}\mathbf{W}_x^T \left(\mathbf{W}_x\mathbf{W}_x^T + \epsilon m\mathbf{G}^{\psi}_{\mathcal{Y}\mathcal{Y}}{}^{-1}\right)^{-1}\mathbf{G}^k_{\mathcal{Y}\mathcal{Y}}{}^{-1} \tag{30}$$

This is computationally advantageous as with $|\mathcal{Y}| = \hat{m} \ll m$ the complexity reduces to $\mathcal{O}(\hat{m}^3 + \hat{m}^2 m)$ and $\mathcal{O}(\hat{m}^2)$ for required pre-computations and per iteration respectively, often with minimal effects on the obtained results.

B.2 Importance Sampling

The estimator (18) is based on a sample from the distribution $p_{\pi^0}(X'|X)\mu(X)$ and while we are free to choose μ, p_{π^0} is specified by $X_{i+1}|X_i$, i.e. the uncontrolled dynamics. In practice

it may be impractical to sample according to the uncontrolled dynamics, e.g., we may wish to improve the policy sequentially collecting new sample following the already learned, rather then the uninformed, policy. To address such situation we follow the importance sampling approach. Specifically note that

$$\mathcal{C}_{ZY}^{kl} = \mathbb{E}_{(Z',Y')} \left[\frac{P(Z',Y')}{Q(Z',Y')} k(Z,\cdot) \otimes l(Y,\cdot) \right] \ , \tag{31}$$

where P, Q are the p.d.f.s of the two joints (Z, Y), (Z', Y') and we assume $Q(z, y) = 0 \Rightarrow P(z, y) = 0$. Hence given a i.i.d. sample from (Z', Y') and empirical estimate of $\mathcal{C}_{ZY}^{kl}$ is given by

$$\hat{\mathcal{C}}_{\mathcal{D}}^{kl} = \sum_{i=1}^{m} w_i k(\cdot, z_i) \otimes l(\cdot, y_i) \quad , \text{with} \quad w_i = P(z_i, y_i)/Q(z_i, y_i) \ . \tag{32}$$

Applying these to (11) to obtain an empirical estimate of $\mathcal{U}$, it is easy to show that the based on a sample from $p_\pi(X'|X)\mu(X)$, formed from an alternative policy, the estimator $\bar{\Psi}_i = \mathbf{g}_{\mathcal{X}}^\psi \alpha_i$ with

$$\alpha_i = \left[\mathbf{G}_{\mathcal{DB}}^\phi \beta \odot \mathbf{G}_{\mathcal{X}'\mathcal{A}}^\psi \alpha_{i+1} \right]^T \mathbf{W} (\mathbf{G}_{\mathcal{X}\mathcal{X}}^k + \epsilon n \mathbf{I})^{-1} \tag{33}$$

is obtained, where $\mathbf{W}$ is the diagonal weight matrix with $\mathbf{W}_{ii} = p_{\pi^0}(x_i'|x_i)/p_\pi(x_i'|x_i)$ and we again assume that $p_\pi(x'|x) = 0 \Rightarrow p_{\pi^0}(x'|x) = 0$.

C Proofs and Derivations

C.1 General Results for Path Integral Control

Theorem 1. *Let the optimal value function be bounded, say $J^*(\cdot, t) < c$ then,*

$$\|\bar{\Psi}(\cdot, t) - \Psi(\cdot, t)\|_\infty \to 0 \implies \|\bar{J}^*(\cdot, t) - J^*(\cdot, t)\|_\infty \to 0 \tag{34}$$

Proof. From (3),

$$J^*(\cdot, t) < c \Rightarrow \Psi(\cdot, t) > c' > 0 \tag{35}$$

Now

$$\|\bar{J}^*(\cdot, t) - J^*(\cdot, t)\|_\infty = \sup_{\mathbf{x}} |\bar{J}^*(\mathbf{x}, t) - J^*(\mathbf{x}, t)| \tag{36}$$

$$= \lambda \sup_{\mathbf{x}} |\log \frac{\bar{\Psi}(\mathbf{x}, t)}{\Psi(\mathbf{x}, t)}| \tag{37}$$

$$= \lambda \sup_{\mathbf{x}} |\log(\frac{\bar{\Psi}(\mathbf{x}, t) - \Psi(\mathbf{x}, t)}{\Psi(\mathbf{x}, t)} + 1)| \tag{38}$$

$$\leq \lambda \sup_{\mathbf{x}} |\log(\frac{\bar{\Psi}(\mathbf{x}, t) - \Psi(\mathbf{x}, t)}{c'} + 1)| \tag{39}$$

and thus

$$\|\bar{\Psi}(\cdot, t) - \Psi(\cdot, t)\|_\infty \to 0 \Rightarrow \frac{\bar{\Psi}(\mathbf{x}, t) - \Psi(\mathbf{x}, t)}{c'} + 1 \to 1 \Rightarrow \|\bar{J}^*(\cdot, t) - J^*(\cdot, t)\|_\infty \to 0 \tag{40}$$

□

C.2 Convergence of Estimates

Theorem 2. *Under the assumptions in the main text, the assumptions of lemma 3 and assuming all relevant kernels satisfy $0 \leq k(x, x') \leq 1$, the estimator $\bar{\Psi}_i$ is consistent, i.e., $\|\bar{\Psi}_i - \Psi_i\|_{\mathcal{H}}$ converges in probability.*

Proof. Let $\tilde{\Psi}_i = \hat{\mathcal{U}}^* [\Phi \otimes \Psi_{i+1}]$ where $\hat{\mathcal{U}}^*$ is the adjoint of $\hat{\mathcal{U}}$, i.e., $\tilde{\Psi}_i$ captures the approximation arising due to the empirical embedding. Then using general relation $\|\mathcal{T}h\|_{\mathcal{H}} \leq \|\mathcal{T}\|_2 \|h\|_{\mathcal{H}} \leq \|\mathcal{T}\|_{HS} \|h\|_{\mathcal{H}}$ for bounds on operators, we have the bound

$$\|\tilde{\Psi}_i - \Psi_i\|_{\mathcal{H}} = \|\hat{\mathcal{U}}^* [\Phi \otimes \Psi_{i+1}] - \mathcal{U}^* [\Phi \otimes \Psi_{i+1}]\|_{\mathcal{H}} \tag{41}$$

$$\leq \|\Phi \otimes \Psi_{i+1}\|_{\mathcal{H}} \|\hat{\mathcal{U}}^* - \mathcal{U}^*\|_{HS} \tag{42}$$

$$= \|\Phi\|_{\mathcal{H}} \underbrace{\|\Psi_{i+1}\|_{\mathcal{H}} \|\hat{\mathcal{U}}^* - \mathcal{U}^*\|_{HS}}_{=: \epsilon_i} \tag{43}$$

Now

$$\|\bar{\Psi}_i - \Psi_i\|_{\mathcal{H}} \leq \|\bar{\Psi}_i - \tilde{\Psi}_i\|_{\mathcal{H}} + \|\tilde{\Psi}_i - \Psi_i\|_{\mathcal{H}} \tag{44}$$

$$\leq \|\hat{\mathcal{U}}^* [\Phi \otimes \bar{\Psi}_{i+1}] - \hat{\mathcal{U}}^* [\Phi \otimes \Psi_{i+1}]\|_{\mathcal{H}} + \|\Phi\|_{\mathcal{H}} \epsilon_i \tag{45}$$

$$\leq \|\Phi \otimes \bar{\Psi}_{i+1} - \Phi \otimes \Psi_{i+1}\|_{\mathcal{H}} \|\hat{\mathcal{U}}^*\|_2 + \|\Phi\|_{\mathcal{H}} \epsilon_i \tag{46}$$

$$\leq \|\Phi\|_{\mathcal{H}} \|\bar{\Psi}_{i+1} - \Psi_{i+1}\|_{\mathcal{H}} + \|\Phi\|_{\mathcal{H}} \epsilon_i \tag{47}$$

where in the last line we used $0 \leq k(x, x') \leq 1 \Rightarrow \|\hat{\mathcal{U}}^*\|_2 \leq 1$. As we may further using lemma 3 and the union bound construct ϵ s.t. with probaility $1 - \delta$ simultaniously for all ϵ_i, $\epsilon_i \leq \epsilon$. The result then follows by induction. $\qquad \square$

C.3 Auxillary Results

Lemma 3 (Song et al., 2010)**.** *Assume the operator $\mathcal{C}_{YX} \mathcal{C}_{XX}^{-\frac{3}{2}}$ is Hilbert-Schmidt, then*

$$\|\hat{\mathcal{U}} - \mathcal{U}_{Y|X}\|_{HS} = \mathcal{O}(\lambda^{\frac{1}{2}} + \lambda^{-\frac{3}{2}} m^{-\frac{1}{2}}) \tag{48}$$

In particular if the regularization term λ satisfies $\lambda \to 0$ and $m\lambda^3 \to \infty$, then $\|\hat{\mathcal{U}}_{\mathcal{D}}^{lk} - \mathcal{U}^{lk}\|_{HS}$ converges in probability.

Proof. See [13] Theorem 1. $\qquad \square$