Reproducing kernel Hilbert space semantics for probabilistic programs

Adam Ścibior
University of Cambridge
and MPI Tübingen

Bernhard Schölkopf
MPI Tübingen

Abstract

We propose denotational semantics for a language of probabilistic arithmetic expressions based on reproducing kernel Hilbert spaces (RKHS). The RKHS approach has numerous practical advantages, but from a semantics point of view the most important is ability to provide convergence guarantees on approximate evaluations of expressions. We present preliminary results on convergence bounds, adapting them to more general settings is still work in progress.

1. A grammar of probabilistic expressions

As an example of a probabilistic programming language we consider a simple grammar of probabilistic expressions.

$$\begin{aligned} e ::= \quad & r \quad | \quad v \quad | \quad f(e_1, e_2) \quad | \quad D(e_1, e_2) \\ | \quad & \textbf{let}\ v = e_1 \textbf{in}\ e \end{aligned}$$

where r is a real number, v is a variable, f is a deterministic (measurable) function from a predefined collection, D is a probability distribution (parametrised by real numbers) from a predefined collection. Although for simplicity we only consider real-valued primitives here, we emphasize that our approach is equally applicable to other data types, such as integers or strings, as long as we can define a positive definite kernel for them. Similarly, the syntax only includes binary functions for clarity of notation, but functions of arbitrary arity could be used instead.

Expressions generated by this grammar correspond to probability distributions over the set of real numbers, as long as they contain no free variables. It is straightforward to give semantics to those expressions using e.g. the probability monad, but in practice the required computation may be prohibitively expensive. We show how to derive equivalent semantics based on RKHS and how to compute it approximately, potentially with convergence guarantees.

2. Introduction to RKHS

Our approach is based on the existing large body of work on RKHS (Berlinet and Thomas-Agnan 2004; Schölkopf and Smola 2001). This is a vast topic, so we only provide a short introduction to the main concepts below.

The basic idea is to map points in the input space $\mathscr{X}$ (here $\mathscr{X} = \mathbb{R}$) to a feature space $\mathscr{H}$ where the relationships we are interested in have a simpler algebraic form. For example, in support vector machines (SVM) non-linear boundaries in the input space become linear in the feature space. However, the feature space is usually higher dimensional than the input space, so performing computation in it is more expensive.

Working explicitly with elements of the feature space can be avoided if the feature space $\mathscr{H}$ is an RKHS. To construct an RKHS we start with a symmetric function $k : \mathscr{X} \times \mathscr{X} \to \mathbb{R}$ satisfying the following condition: for any $m \in \mathbb{N}$, $a_1, \ldots, a_m \in \mathbb{R}$, and $x_1, \ldots, x_m \in \mathscr{X}$

$$\sum_{i,j=1}^{m} a_i a_j k(x_i, x_j) \geq 0.$$

Such a function is called a positive definite kernel.

We say that $\mathscr{H}$ is an RKHS for k if there exists a mapping $\Phi : \mathscr{X} \to \mathscr{H}$ such that $\langle \Phi(x), \Phi(x') \rangle = k(x, x')$. For our purposes the exact structure of $\mathscr{H}$ does not matter, as long as this relationship holds. We should note, however, that for every positive definite k it is possible to construct such $\mathscr{H}$ and Φ.

We can extend the mapping Φ to distributions over $\mathscr{X}$ (Smola et al. 2007). Let $\mathscr{P}_+^1(\mathscr{X})$ denote the set of all Borel probability measures over set $\mathscr{X}$. For $p \in \mathscr{P}_+^1(\mathscr{X})$ satisfying

$$\mathbb{E}_{x,x' \sim p}[k(x, x')] < \infty$$

we can define

$$\begin{aligned} \mu &: \mathscr{P}_+^1(\mathscr{X}) \to \mathscr{H} \\ \mu(p) &= \mathbb{E}_{x \sim p}[\Phi(x)] \end{aligned}$$

which allows us to map probability distributions to elements of $\mathscr{H}$. From now on we assume that k is a characteristic kernel, which means that μ is injective (Steinwart 2002). Apart from that the choice of kernel is largely arbitrary, although in practice it has a large impact on the quality of results when only a finite number of samples is used. In machine learning two popular kernels are Gaussian

$$k(x, x') = \exp\left(-\frac{\|x - x'\|^2}{2\sigma^2}\right)$$

and Laplacian

$$k(x, x') = \exp\left(-\frac{\|x - x'\|}{\sigma}\right)$$

.

3. Kernel-based semantics

We now turn to defining semantics of the grammar of probabilistic expressions above in terms of RKHS elements. We fix the RKHS $\mathscr{H}$ by selecting a particular kernel k. By construction $\mathscr{H}$ is separable, which implies that any $h \in \mathscr{H}$ can be written as a potentially infinite sum of the form $\sum_i \alpha_i \Phi(x_i)$, where for all i $\alpha_i \in \mathbb{R}$, $x_i \in \mathscr{X}$. We use this property for writing the denotations of subexpressions as

$$\llbracket e_1 \rrbracket_k = \sum_i \alpha_i \Phi(x_i)$$
$$\llbracket e_2 \rrbracket_k = \sum_j \beta_j \Phi(y_j)$$

Using those expansions we can write the semantics as

$$\begin{aligned}
\llbracket r \rrbracket_k &= \Phi(r) \\
\llbracket f(e_1, e_2) \rrbracket_k &= \sum_{i,j} \alpha_i \beta_j \Phi(f(x_i, y_j)) \\
\llbracket D(e_1, e_2) \rrbracket_k &= \sum_{i,j} \alpha_i \beta_j \mu(D(x_i, y_j)) \\
\llbracket \textbf{let}\, v = e_1 \textbf{in}\, e \rrbracket_k &= \sum_i \alpha_i \llbracket e[x_i/v] \rrbracket_k
\end{aligned}$$

If k is a characteristic kernel this semantics is equivalent to a standard one based on measure theory. To be precise, if we write $P(\cdot)$ for the standard semantics, then for any e we have

$$\begin{aligned}
\llbracket e \rrbracket_k &= \mu(P(e)) \\
\mu^{-1}(\llbracket e \rrbracket_k) &= P(e)
\end{aligned}$$

However, the RKHS representation can be more convenient for further processing. We list several applications of RKHS in section 5.

Of course, obtaining exact results is in general intractable. In the RKHS formulation this is witnessed by the potentially infinite sums in the definition of semantics. In practice we need to truncate the sums to keep the computational cost feasible. Choosing a suitable truncation is a task well-studied in the RKHS literature, usually referred to as ”reduced expansion set methods”. From a theoretical perspective, however, it is more interesting to note that under certain conditions it is possible to provide concrete convergence guarantees for the approximate expansions (Sriperumbudur et al. 2008).

4. Convergence guarantees

This section is work in progress and we only present preliminary results that indicate what kinds of guarantees can be possible. We are working on generalising those results to a setting such as the semantics above.

The first theorem (Schölkopf et al. 2015, theorem 2) applies to situations where we compute functions of random variables for which we can generate perfect samples from their distributions.

Theorem 1. *Let X and Y be independent random variables, with i.i.d. samples $x_1, \dots, x_m$ and $y_1, \dots, y_m$ accordingly. For any measurable function f we have*

$$\left\| \frac{1}{m^2} \sum_{i,j=1}^{m} \Phi(f(x_i, y_j)) - \mu[f(X,Y)] \right\| = O_p\left(\frac{1}{\sqrt{m}}\right)$$

as $m \to \infty$.

The second theorem (Schölkopf et al. 2015, theorem 3) applies when computing functions of RKHS approximations, where the samples may not be uniformly weighted. However, it still assumes that the samples are independent and identically distributed (i.i.d.), which is not generally the case for the semantics above.

Theorem 2. *Let X and Y be independent random variables, with i.i.d. samples $x_1, \dots, x_m$ and $y_1, \dots, y_m$ accordingly. Assume the constants $(\alpha_i)_{i=1}^m$ and $(\beta_i)_{i=1}^m$ satisfy $\sum_{i=1}^m \alpha_i = 1$ and $\sum_{i=1}^m \beta_i = 1$. Assume $\lim_{m\to\infty} \sum_{i=1}^m \alpha_i^2 = \lim_{m\to\infty} \sum_{i=1}^m \beta_i^2 = 0$. Then*

$$\left\| \sum_{i,j=1}^{m} \alpha_i \beta_j \Phi(f(x_i, y_j)) - \mu[f(X,Y)] \right\| = O_p\left(\sqrt{\sum_i \alpha_i^2} + \sqrt{\sum_i \beta_i^2} \right)$$

as $m \to \infty$.

Theorem 1 is a special case of theorem 2.

5. Ongoing work

We are currently exploring two main directions extending this approach. The first is to provide compositional convergence guarantees, similar to those shown in section 4, ideally covering the entire semantics presented in section 3. This would in particular require relaxing the i.i.d. assumption from theorem 2.

Another direction is extending the semantics presented above to deal with more general models and conditioning. For the latter we hope to employ conditional mean embedding operators (Song et al. 2009), but we do not have a clear idea how to approach the former.

Finally, an important question is what to do with the result expressed as an RKHS element. It could either be converted to a probability measure, at a significant computational cost, or used in another method that takes RKHS elements as input, such as kernel two-sample testing (Gretton et al. 2012) or kernel independent component analysis (Bach and Jordan 2003).

References

F. Bach and M. Jordan. Kernel independent component analysis. *J. Mach. Learn. Res.*, 3:1–48, 2003.

A. Berlinet and C. Thomas-Agnan. *Reproducing kernel Hilbert spaces in probability and statistics*. Kluwer Academic, 2004. ISBN 1-4020-7679-7.

A. Gretton, K. M. Borgwardt, M. J. Rasch, B. Schölkopf, and A. Smola. A kernel two-sample test. *J. Mach. Learn. Res.*, 13:723–773, Mar. 2012. ISSN 1532-4435.

B. Schölkopf and A. J. Smola. *Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond*. MIT Press, Cambridge, MA, USA, 2001. ISBN 0262194759.

B. Schölkopf, K. Muandet, K. Fukumizu, S. Harmeling, and J. Peters. Computing functions of random variables via reproducing kernel Hilbert space representations. *Statistics and Computing*, 25(4):755–766, 2015. ISSN 0960-3174. .

A. Smola, A. Gretton, L. Song, and B. Schölkopf. A Hilbert space embedding for distributions. In M. Hutter, R. Servedio, and E. Takimoto, editors, *Algorithmic Learning Theory*, volume 4754 of *Lecture Notes in Computer Science*, pages 13–31. Springer Berlin Heidelberg, 2007. ISBN 978-3-540-75224-0. .

L. Song, J. Huang, A. Smola, and K. Fukumizu. Hilbert space embeddings of conditional distributions with applications to dynamical systems. In *Proceedings of the 26th Annual International Conference on Machine Learning*, pages 961–968, 2009.

B. K. Sriperumbudur, A. Gretton, K. Fukumizu, G. Lanckriet, and B. Schölkopf. Injective Hilbert space embeddings of probability measures. In *COLT*, 2008.

I. Steinwart. On the influence of the kernel on the consistency of support vector machines. *J. Mach. Learn. Res.*, 2:67–93, Mar. 2002. ISSN 1532-4435. .