Nonparametric Tree Graphical Models via Kernel Embeddings

Le Song,[1] **Arthur Gretton,**[1,2] **Carlos Guestrin**[1]
[1] School of Computer Science, Carnegie Mellon University; [2]MPI for Biological Cybernetics

Abstract

We introduce a nonparametric representation for graphical model on trees which expresses marginals as Hilbert space embeddings and conditionals as embedding operators. This formulation allows us to define a graphical model solely on the basis of the feature space representation of its variables. Thus, this nonparametric model can be applied to general domains where kernels are defined, handling challenging cases such as discrete variables with huge domains, or very complex, non-Gaussian continuous distributions. We also derive *kernel belief propagation*, a Hilbert-space algorithm for performing inference in our model. We show that our method outperforms state-of-the-art techniques in a cross-lingual document retrieval task and a camera rotation estimation problem.

1 Introduction

Probabilistic graphical models have become a key tool for representing structured dependencies between random variables in challenging tasks in social networks, computational biology, natural language processing, computer vision, and beyond. Unfortunately, most successful applications of graphical models rely on situations where each random variable can take on only a relatively small number of values, or, in continuous domains, where their joint distributions are Gaussians. In this paper, we present a novel nonparametric representation for tree-structured graphical models that allows us to concisely represent distributions and conduct inference in highly non-Gaussian continuous settings, and in discrete settings where variables can take on a huge number of assignments.

Appearing in Proceedings of the 13^{th} International Conference on Artificial Intelligence and Statistics (AISTATS) 2010, Chia Laguna Resort, Sardinia, Italy. Volume 9 of JMLR: W&CP 9.

At the core of our approach, we exploit the success of kernel methods, which are able to determine nonlinear relations and make predictions from complex and structured data (such as documents, strings, and images; see Schölkopf et al., 2004). These results have been achieved by applying linear techniques on data mapped to a reproducing kernel Hilbert space (RKHS). To date, probabilistic models using RKHS approaches have generally been very simple, with just two variables connected by an edge (e.g., the inputs X and labels Y, drawn *i.i.d.* from a distribution $\mathbb{P}_{XY}$). In this paper, we are able to apply such kernel-based techniques to any tree-structured graphical model.

In the context of graphical models, kernels have thus far been used to represent complex structure in observed variables, but the relationship between unobserved variables has remained simple (c.f., Taskar et al., 2004). In contrast, we define a new representation for graphical model on trees, based on RKHS embeddings of both the node marginals and the conditionals associated with edges between variables. We make use of recent work in formulating embeddings (Berlinet & Thomas-Agnan, 2003; Gretton et al., 2007; Smola et al., 2007; Sriperumbudur et al., 2008) and conditional embeddings (Song et al., 2009) of probabilities into reproducing kernel Hilbert spaces. These embeddings characterize the probabilities solely on the basis of their feature space representations, thus allowing us to deal easily with distributions on variables with complex structure and in high dimensions.

Inference (for instance, determining conditional probabilities and marginals) is key to making predictions using graphical models, and may be conducted using messages passed between the variables along the edges. Such message passing approaches work well when each node can take a small and finite set of values, or where the distribution conforms to a known parametric model (e.g., jointly Gaussian, in the case of Weiss & Freeman, 2001). More recently, there has been a focus on formulating more flexible message passing algorithms on graphical models, however these rely on either a mixture model (Sudderth et al., 2003) or a

set of samples (Ihler & McAllester, 2009) to represent the messages, and thus do not easily generalize to high dimensions or structured data.

In addition to our kernel representation of the node marginals and conditionals, we also derive *kernel belief propagation*, a Hilbert space version of the belief propagation algorithm (Pearl, 1988) required for efficient inference. Unlike Ihler & McAllester (2009), for example, who assume that the node and edge potentials are given in advance and that they can sample from certain distributions associated with these potentials, we learn the model representation from data, do not require sampling in the inference procedure, and study the sample complexity of inference based on models we learn. Our experiments on two domains, cross-language document retrieval and learning camera orientation from images, demonstrate significant improvements of our methods over state-of-the-art techniques.

2 Inference on trees

We begin by introducing undirected tree graphical models, and describe the elimination algorithm for inferring beliefs on the marginals. Our new approach, *Kernel Belief Propagation*, will later be used to express this algorithm in terms of RKHS covariance operators.

2.1 Tree graphical models

Let $\mathcal{T}$ be a tree consisting of nodes $\mathcal{V} = \{1, \ldots, n\}$ and edges $\mathcal{E}$, and denote by $\Gamma_s = \{t|(s,t) \in \mathcal{E}\}$ the set of neighbors of node s in $\mathcal{T}$. In a probabilistic graphical model each node $s \in \mathcal{V}$ is associated with a random variable X_s taking on values in domain $\mathcal{X}_s$ This model can be compactly represented as a set of marginals $\mathbb{P}(X_s)$, one for each node s, and a set of joints $\mathbb{P}(X_s, X_t)$, each associated with an edge (s,t). The joint distribution for all nodes is

$$\mathbb{P}(\{X_s\}_{s=1}^n) = \prod_{(s,t)\in\mathcal{E}} \mathbb{P}(X_s, X_t) \prod_{s\in\mathcal{V}} \mathbb{P}(X_s)^{1-d_s},$$

where $d_s := |\Gamma_s|$ is the cardinality of Γ_s. An interesting property of this undirected tree graphical model is that one can always re-root the tree, and express $\mathbb{P}(\{X_s\}_{s=1}^d)$ using the root marginal $\mathbb{P}(X_r)$ and a set of conditionals $\mathbb{P}(X_s|X_{\pi(s)})$ corresponding to directed edges from node $\pi(s)$ to node s,

$$\mathbb{P}(\{X_s\}_{s=1}^n) = \mathbb{P}(X_r) \prod_{s\in\mathcal{V}\setminus t} \mathbb{P}(X_s|X_{\pi(s)}), \qquad (1)$$

where we choose node r as the root and denote the parent of a node s by $\pi(s)$. Given evidence $\bar{x}_t$ in a set $\mathcal{L} \subseteq \mathcal{V} \setminus r$ of leaf nodes, and a set $\mathcal{I} := \mathcal{V} \setminus \mathcal{L} \setminus r$ of internal nodes, the marginal of the root node r is

$$\mathbb{P}(X_r, \{\bar{x}_t\}_{t\in\mathcal{L}}) = \qquad (2)$$
$$\int_{X_s\in\mathcal{X}, s\in\mathcal{I}} \mathbb{P}(X_r) \prod_{t\in\mathcal{L}} \mathbb{P}(\bar{x}_t|X_{\pi(t)}) \prod_{s\in\mathcal{I}} \mathbb{P}(X_s|X_{\pi(s)}).$$

If a particular piece of evidence is not on a leaf node, then all of the descendants of this node (both observed and unobserved) can be pruned for the inference step. After pruning, this node will now be a leaf, and the result of inference will be unchanged.

2.2 Belief propagation

The expression (2) can be computed efficiently by belief propagation (Pearl, 1988);(Jordan, 2002, Ch. 3,4). This is done by passing messages m_{ts} from nodes t to s, starting from the leaves and progressing up to the root. Messages to the target node s are functions on the state space of $\mathcal{X}_s$, and are defined recursively,

$$m_{ts}(x_s) = \mathbb{E}_{X_t|x_s}\left[\prod_{u\in\Gamma_t\setminus s} m_{ut}(x_t)\right]$$
$$= \mathbb{E}_{X_t|x_s}[M_{ts}(X_t)] \qquad (3)$$

where $M_{ts}(X_t) := \prod_{u\in\Gamma_t\setminus s} m_{ut}(X_t)$ is the "pre-message". If t is a leaf node where a particular observation $\bar{x}_t$ is provided, then it simply sends the message $m_{ts}(x_s) = \mathbb{P}(\bar{x}_t|x_s)$. The belief of the root node X_r is:

$$B_r(x_r) = \mathbb{P}(x_r) \prod_{s\in\Gamma_r} m_{sr}(x_r) = \mathbb{P}(x_r) M_{rs}(x_r) m_{sr}(x_r) \qquad (4)$$

for any $s \in \Gamma_r$. Note that m_{ts} may be normalized to a unit sum for numerical stability, but this is not required for the analysis. In the case where the domains $\mathcal{X}$ is discrete with small cardinality $|\mathcal{X}|$, or $\mathcal{X}$ is continuous but the random variables are Gaussians (Weiss & Freeman, 2001), computing the above marginal can be carried out efficiently using the sum-product algorithm. For general continuous domains or discrete domains where $|\mathcal{X}|$ is too large to enumerate, however, the expectation in (3) becomes intractable.

A number of approaches have been used to define belief propagation in higher dimensional spaces, and for more complex probability models. Minka (2001) proposes the expectation-propagation algorithm, where only certain moments of the messages are estimated. Unfortunately, this method does not address distributions that cannot be well-characterized by the first few moments. To address this problem, Sudderth et al. (2003) represent messages as mixtures of Gaussians, however the number of mixture components grows exponentially as the message is propagated: they alleviate this problem through subsampling. Ihler & McAllester (2009) propose a particle BP approach,

where messages are expressed as functions of a distribution of particles at each node, and the expectation in (3) becomes sums over the particles. In this work, however, the marginals $\mathbb{P}(X_s)$ and joint distribution $\mathbb{P}(X_s, X_t)$ are both assumed known. Furthermore, Ihler & McAllester assume the ability to efficiently obtain samples from the space of outgoing messages given the incoming message, which is not viable for very complex distributions.

The key idea of our algorithm is that we use RKHS functions to express the messages $m_{ts}(x_s)$ between pairs of nodes. As a result of this representation, the messages $m_{ut}(x_t)$ from nodes $u \in \Gamma_t \setminus s$ can be combined in a straightforward linear operation in feature space that implements the sum and product steps, producing a new message $m_{ts}(x_s)$ that remains an RKHS function. Being defined in RKHSs, our method has no difficulty in propagating messages on structured objects or in high dimensions. Moreover, unlike Ihler & McAllester (2009), we are not given the marginals $\mathbb{P}(X_s)$ and joint distribution $\mathbb{P}(X_s, X_t)$ in advance and we do not need to perform sampling. Rather, all functions needed for the algorithm are learned non-parametrically from data, and thus can represent very complex, high-dimensional probability models efficiently. We provide convergence guarantees on the beliefs of the nodes, demonstrating that our approach converges to the true beliefs in the limit of increasing samples used to learn the RKHS representations of the conditional probabilities.

3 Hilbert space embeddings

In the present section, we define basic terms from RKHS theory, and introduce some necessary concepts from functional analysis. In our inference procedure on trees, we will represent both the messages and the beliefs as functions in Hilbert spaces. Let $\mathcal{F}$ be an RKHS on the separable[1] metric space $\mathcal{X}$, with a continuous feature mapping $\varphi(x) \in \mathcal{F}$ for each $x \in \mathcal{X}$. The inner product between feature mappings is given by the positive definite kernel function $k(x, x') := \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}}$, and for all $f \in \mathcal{F}$ and $x \in \mathcal{X}$ we have $\langle f, \varphi(x) \rangle_{\mathcal{F}} = f(x)$. Examples of kernels include the Gaussian RBF kernel $k(x, x') = \exp(-\sigma \|x - x'\|^2)$, however kernel functions have also been defined on graphs, time series, dynamical systems, images, and other structured objects (Schölkopf et al., 2004).

We next define the embedding of a probability distribution in an RKHS. Let $\mathcal{P}$ be the set of Borel probability measures on $\mathcal{X}$, and define as X the random variable with distribution $\mathbb{P} \in \mathcal{P}$ (samples will be written in lowercase). Following Berlinet & Thomas-Agnan (2003); Fukumizu et al. (2004); Gretton et al. (2007); Smola et al. (2007), we define the mapping to $\mathcal{F}$ of $\mathbb{P} \in \mathcal{P}$ as the expectation of $\varphi(x)$ with respect to $\mathbb{P}$, or $\mu_X := \mathbb{E}_X(\varphi(X))$. We refer to μ_X as the mean map. For all $f \in \mathcal{F}$, we may write $\mathbb{E}_{X \sim \mathbb{P}} f(X) = \langle f, \mu_X \rangle_{\mathcal{F}}$. We may think of the mean map by analogy with a mean vector in a finite dimensional space: if $\mathcal{F} = \mathbb{R}^d$, then $f \in \mathbb{R}^d$ is some fixed vector, X is a random vector defined on $\mathbb{R}^d$ with mean[2] μ_X, and $\mathbb{E}_{X \sim \mathbb{P}} \langle f, X \rangle_{\mathcal{F}} = f^\top \mu_X$. A characteristic RKHS is one for which the mean map is injective: that is, distributions have a unique embedding (Sriperumbudur et al., 2008). This property holds for many commonly used kernels on $\mathbb{R}^d$, including the Gaussian, Laplace, and B-spline kernels.

Table 1: Table of Notation

random variable	X	Y
domain	$\mathcal{X}$	$\mathcal{Y}$
observation	x	y
kernel	$k(x, x')$	$l(y, y')$
kernel matrix	K	L
feature map	$\varphi(x), k(x, \cdot)$	$\phi(y), l(y, \cdot)$
feature matrix	Υ	Φ
feature column	$k_x = \Upsilon^\top \varphi(x)$	$l_y = \Phi^\top \phi(y)$
RKHS	$\mathcal{F}$	$\mathcal{G}$

Given a sample $\mathcal{D}_X := \{x_i\}_{i=1}^m$, an empirical estimate of the mean map is straightforward: $\hat{\mu}_X := m^{-1} \sum_{i=1}^m \varphi(x_i) = m^{-1} \Upsilon \mathbf{1}_m$, where $\Upsilon := [\varphi(x_1) \dots \varphi(x_m)]$ is an arrangement of feature space mappings into columns (this is a slight abuse of notation, as the mappings may be infinite dimensional), and $\mathbf{1}_m$ is an $m \times 1$ column of ones.

Consider now the case where we have a joint distribution $\mathbb{P}(X, Y)$ over two random variables X on $\mathcal{X}$ and Y on $\mathcal{Y}$. In defining a feature space characterization of the relation between X and Y, we relate the feature space mapping of X to $\mathcal{F}$ with the feature space map of Y to a second RKHS $\mathcal{G}$ on $\mathcal{Y}$ with kernel $l(y, y')$ (see Table 1 for a summary of notation). The idea is to express the dependence between the feature maps of X and Y using a generalization of covariance to feature spaces.

A definition of covariance between RKHSs requires a generalization of the outer product, since if $x \in \mathbb{R}^d$ and $y \in \mathbb{R}^{d'}$, the (uncentered) covariance matrix between random vectors X and Y is $C_{XY} := \mathbb{E}(X Y^\top)$. We thus introduce the rank one operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ such that

$$f \otimes g(h) = \langle g, h \rangle_{\mathcal{G}} f. \tag{5}$$

[1] A Hilbert space is separable iff it has a countable orthonormal basis, which will be needed when defining singular value decompositions of infinite dimensional operators.

[2] There is a subtlety in that the mean vector must be shown to exist when the feature space is infinite dimensional: see (Sriperumbudur et al., 2008, Theorem 3).

This can be understood by analogy with the finite dimensional case: if $x \in \mathbb{R}^d$ and y, z are vectors in $\mathbb{R}^{d'}$, then $(x\, y^\top)z = x(y^\top z)$. Following (Baker, 1973; Fukumizu et al., 2004), we define the uncentered *covariance operator* $\mathcal{C}_{XX} : \mathcal{F} \to \mathcal{F}$ such that for all $f \in \mathcal{F}$,

$$\begin{aligned}\langle f, \mathcal{C}_{XX} f\rangle_{\mathcal{F}} &= \mathbb{E}_X \langle f, \varphi(X) \otimes \varphi(X) f\rangle_{\mathcal{F}} \\ &= \mathbb{E}_X \left[(f(X))^2\right];\end{aligned}$$

and the uncentered cross-covariance operator $\mathcal{C}_{XY} : \mathcal{G} \to \mathcal{F}$, where for all $f \in \mathcal{F}$ and $g \in \mathcal{G}$,

$$\langle f, \mathcal{C}_{XY} g\rangle_{\mathcal{F}} := \mathbb{E}_{XY}\left[f(X)g(Y)\right]$$

Empirical estimates may again be obtained by substituting sample averages: for example,

$$\left\langle f, \hat{\mathcal{C}}_{XY} g\right\rangle_{\mathcal{F}} = \frac{1}{m}\sum_{i=1}^m f(x_i)g(y_i).$$

4 Conditional Embeddings

We now provide two main results: first, we review the concept of the conditional distribution embedding from Song et al. (2009), although with the addition of a new convergence result (Theorem 1). We then apply Bayes' law to express the likelihood in terms of RKHS covariance operators. These two results will form the basis for our belief propagation algorithm in Sec. 5.

4.1 Embedding conditional distributions

By analogy with the mean embedding in an RKHS, Song et al. represent the conditional distribution $\mathbb{P}(Y|x)$ via a conditional embedding $\mu_{Y|x}$, such that for all $g \in \mathcal{G}$,

$$\left\langle g, \mu_{Y|x}\right\rangle_{\mathcal{G}} = \mathbb{E}_{Y|x}[g(Y)].$$

A case of particular interest is when g is a Parzen window with fixed bandwidth, for which case we obtain a (smoothed) conditional density estimate. We emphasize that $\mu_{Y|x}$ now traces out a set of mean embeddings in $\mathcal{G}$, with each element corresponding to a particular value of x. The conditional mean mapping is defined via a conditional embedding operator $\mathcal{U}_{Y|X} : \mathcal{F} \mapsto \mathcal{G}$

$$\mu_{Y|x} = \mathcal{U}_{Y|X}\varphi(x) = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\varphi(x). \tag{6}$$

Given dataset $\mathcal{D}_{XY} = \{(x_i, y_i)\}_{i=1}^m$ drawn *i.i.d.* from $\mathbb{P}(X, Y)$, the conditional embedding operator $\mathcal{U}_{Y|X}$ has the finite sample estimate

$$\hat{\mathcal{U}}_{Y|X} = \tfrac{\Phi\Upsilon^\top}{m}(\tfrac{\Upsilon\Upsilon^\top}{m} + \lambda I)^{-1} = \Phi(K + \lambda m I)^{-1}\Upsilon^\top, \tag{7}$$

where (with some regrettable notation) we have defined the kernel matrix $K := \Upsilon^\top\Upsilon$ with (i, j)th entry $k(x_i, x_j)$, and the columns of Φ contain the mappings $\phi(y_i)$. Note that we have added an additional regularizing term λ, to avoid overfitting (in much the same way as is done for kernel canonical correlation analysis: see Fukumizu et al., 2007). A novel result of the present work is the consistency of this estimator, which is proved in the appendix.

Theorem 1 *Assume $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ is Hilbert-Schmidt.*[3] *Then $\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\|_{HS} = O_p(\lambda^{\frac{1}{2}} + \lambda^{-\frac{3}{2}}m^{-\frac{1}{2}})$. If the regularization term λ satisfies $\lambda \to 0$ and $m\lambda^3 \to \infty$, then $\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\|_{HS}$ converges in probability.*

Writing $k_x := \Upsilon^\top\varphi(x)$, then $\hat{\mu}_{Y|x}$ becomes $\hat{\mathcal{U}}_{Y|X}\phi(x) = \Phi(K + \lambda m I)^{-1}k_x = \sum_{i=1}^m \beta_x(y_i)\phi(y_i)$, with $\beta_x(y_i) \in \mathbb{R}$. This estimator resembles the empirical estimate $\hat{\mu}_X$ of the unconditional mean embedding, but with the uniform weights $\frac{1}{m}$ replaced by non-uniform weights $\beta_x(y_i)$.

4.2 Bayes rule via embeddings

In the previous section, we used an RKHS function $\mu_{Y|x} \in \mathcal{G}$ to represent $\mathbb{P}(Y|x)$, the posterior distribution of Y given some observed x. For the purpose of conducting inference on graphs, we will also need to represent the likelihood $\mathbb{P}(x|Y)$ of x given Y. We now present a kernel-based estimate of this quantity, under the assumption $\mathbb{P}(x|Y) \in \mathcal{G}$. Consider Bayes' law,

$$\mathbb{P}(Y|x) = \mathbb{P}(x|Y)\mathbb{P}(Y)\left(\mathbb{P}(x)\right)^{-1}.$$

Multiplying both sides by $\phi(Y)$ and integrating over Y yields

$$\mathbb{E}_{Y|x}[\phi(Y)] = \mathbb{E}_Y\left[\phi(Y)\mathbb{P}(x|Y)\right]\left(\mathbb{P}(x)\right)^{-1}.$$

We now define an RKHS function representing the unnormalized likelihood of x given Y, $f_x(\cdot) := \mathbb{P}(x|\cdot)\left(\mathbb{P}(x)\right)^{-1}$, where we assume $f_x(\cdot) \in \mathcal{G}$: under this assumption, $f_x(Y) = \langle f_x, \phi(Y)\rangle_{\mathcal{G}}$. Our assumption follows directly from $\mathbb{P}(x|Y) \in \mathcal{G}$, given $\mathbb{P}(x)$ is constant with respect to Y. We emphasize that while $f_x(\cdot)$ represents the unnormalized likelihood of x, it should be thought of a function in $\mathcal{G}$ *indexed* by x which takes Y as an argument. Note that while $f_x(Y)$ differs from our original goal of $\mathbb{P}(x|Y)$ by a factor $\mathbb{P}(x)$, this is immaterial in the BP algorithm, where the messages being passed need not be normalized probabilities.

To obtain an expression for $f_x(\cdot)$ in terms of covariance operators (and hence an empirical estimate), we replace $\mathbb{E}_{Y|x}[\phi(Y)] = \mathcal{U}_{Y|X}\varphi(x)$, using the conditional embedding operator from the previous section. Then

$$\begin{aligned}\mathcal{U}_{Y|X}\varphi(x) &= \mathbb{E}_Y\left[\langle f_x, \phi(Y)\rangle_{\mathcal{G}}\,\phi(Y)\right] \\ &= \mathbb{E}_Y\left[\phi(Y)\otimes\phi(Y)\right]f_x(\cdot) = \mathcal{C}_{YY}f_x(\cdot),\end{aligned}$$

where in the middle line we used the definition of the tensor product from (5). Defining $\mathcal{A}_{YX} := \mathcal{C}_{YY}^{-1}\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$, we obtain $f_x := \mathcal{A}_{YX}\varphi(x)$.

[3] The Hilbert-Schmidt norm of an operator is the square root of the sum of the squared singular values. A formal definition is given in the appendix.

For finite samples, $\mathcal{A}_{YX}$ has the empirical expression

$$\hat{\mathcal{A}}_{YX} = (\tfrac{\Upsilon\Upsilon^\top}{m} + \lambda I)^{-1} \tfrac{\Upsilon\Phi^\top}{m} (\tfrac{\Phi\Phi^\top}{m} + \lambda I)^{-1} = m\Upsilon((L + \lambda m I)(K + \lambda m I))^{-1}\Phi^\top, \quad (8)$$

where we have used the matrix inversion lemma in the second equality, and $L := \Phi^\top\Phi$ has (i, j)th entry $l(y_i, y_j)$. As with the empirical $\hat{\mathcal{U}}_{Y|X}$, we apply a regularization with scaling λ. Our next theorem, which is again proved in the appendix, shows the empirical estimator to be consistent.

Theorem 2 *Assume $\mathcal{C}_{YY}^{-\frac{3}{2}}\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ is Hilbert-Schmidt. Then $\|\hat{\mathcal{A}}_{YX} - \mathcal{A}_{YX}\|_{HS} = O_p(\lambda^{\frac{1}{2}} + \lambda^{-2}m^{-\frac{1}{2}})$. If $\lambda \to 0$, and $\lambda^4 m \to \infty$, then $\|\hat{\mathcal{A}}_{YX} - \mathcal{A}_{YX}\|_{HS}$ converges in probability.*

The function value $f_x(y)$ can be estimated as $\langle \phi(y), \hat{\mathcal{A}}_{YX}\varphi(x)\rangle_{\mathcal{G}} = mk_x^\top((L + \lambda m I)(K + \lambda m I))^{-1} l_y = \sum_i \gamma_x(y_i) l(y_i, y)$ with $\gamma_x(y_i) \in \mathbb{R}$.

5 BP with RKHS embeddings

We now describe our RKHS version of the belief propagation algorithm in Section 2.2. The key idea is to represent the messages as functions in the RKHS, and propagate the messages via conditional embedding operators. We need three main steps: First, we must define RKHS functions representing the initial messages from the observed leaf nodes. Second, to implement the recursion in (3), we must express a message $m_{ts} \in \mathcal{F}_s$ from t to s in terms of the incoming messages $m_{ut} \in \mathcal{F}_t$ of the remaining neighbours of t, where the subscript on a feature space designates its corresponding node.[4] Third, we must be able to compute the belief at the root node r as a function of all its incoming messages. We now address each of these points in turn.

The kernel expression for the message sent from a leaf node t to a node s is straightforward: recall from Section 2.2 that this message is simply $m_{ts}(x_s) = \mathbb{P}(\bar{x}_t|x_s)$. Consequently, we use the result from Section 4.2 that $\mathbb{P}(\bar{x}_t|\cdot)/\mathbb{P}(\bar{x}_t) = \mathcal{A}_{st}\varphi(\bar{x}_t) \in \mathcal{F}_s$, and define $m_{ts}(\cdot) := \mathcal{A}_{st}\varphi(\bar{x}_t) \in \mathcal{F}_s$ (the scaling by $1/\mathbb{P}(\bar{x}_t)$ is immaterial, as we do not normalize the intermediate messages when running the algorithm). The empirical estimate of m_{ts} is thus given by $\hat{m}_{ts} = \Upsilon_s((L_t + \lambda I)(L_s + \lambda I))^{-1}\Upsilon_t^\top\varphi(\bar{x}_t) = \Upsilon_s\beta_{ts}$, where we have further defined $\beta_{ts} := ((L_t + \lambda I)(L_s + \lambda I))^{-1}\Upsilon_t^\top\varphi(\bar{x}_t)$.

The recursion for internal nodes in the tree is more complex. Recall that

$$m_{ts}(x_s) = \mathbb{E}_{X_t|x_s}\left[\prod\nolimits_{u\in\Gamma_t\setminus s} m_{ut}(X_t)\right].$$

[4] The identity of a feature mapping φ is specified by the node of its argument.

Each of messages $m_{ut}(\cdot)$ is a function in $\mathcal{F}_t$, hence we may write $m_{ut}(X_t) = \langle m_{ut}, \varphi(X_t)\rangle$.

To combine these messages, we begin by defining the notion of inner product between an arbitrary operator A from $\mathcal{G}$ to $\mathcal{F}$, $\langle f \otimes g, A\rangle_{HS} = \langle Ag, f\rangle_{\mathcal{F}}$, (see supplement for the formal definition of this inner product). In particular, from (5), $\langle a \otimes b, u \otimes v\rangle_{HS} = \langle a, u\rangle_{\mathcal{F}} \langle b, v\rangle_{\mathcal{G}}$. Since we now consider the interaction between more than two variables, we must generalize this concept to higher dimensions. Given the tensor product space $\otimes^n\mathcal{F}$ and functions $a_i \in \mathcal{F}$ and $b_i \in \mathcal{F}$ for $i \in \{1, \ldots, n\}$,

$$\left\langle \bigotimes_{i=1}^n a_i, \bigotimes_{i=1}^n b_i \right\rangle_{\otimes^n\mathcal{F}} = \prod_{i=1}^n \langle a_i, b_i\rangle_{\mathcal{F}}. \quad (9)$$

We can then write

$$m_{ts}(x_s) = \mathbb{E}_{X_t|x_s}\left\langle \bigotimes_{u\in\Gamma_t\setminus s} m_{ut}, \bigotimes_{d_t-1} \varphi(X_t)\right\rangle_{\mathcal{H}_t},$$

$d_t := |\Gamma_t|$ is the number of neighbours of t, and $\mathcal{H}_t := \bigotimes_{d_t-1}\mathcal{F}_t$ is the tensor product space made up of $d_t - 1$ of the $\mathcal{F}_t$ spaces. Next, by analogy with (6), we may write

$$\mathbb{E}_{X_t|x_s}\left[\bigotimes\nolimits_{d_t-1}\varphi(X_t)\right] = \mathcal{U}_{X_t^{d_t-1}|X_s}\varphi(x_s),$$

where we have defined the conditional embedding operator $\mathcal{U}_{X_t^{d_t-1}|X_s} : \mathcal{F} \to \mathcal{H}_t$. This operator embeds the conditional distribution $\mathbb{P}(X_t|x_s)$ using a tensor product feature map $\bigotimes_{d_t-1}\varphi(X_t)$.

We also define the pre-message $M_{ts} := \bigotimes_{u\in\Gamma_t\setminus s} m_{ut}$, and obtain

$$m_{ts}(x_s) = \left\langle M_{ts}, \mathcal{U}_{X_t^{d_t-1}|X_s}\varphi(x_s)\right\rangle_{\mathcal{H}_t}.$$

The empirical expression for each incoming message can be written $\hat{m}_{ut} = \Upsilon_t^{(u)}\beta_{ut}$, where $\Upsilon_t^{(u)}$ contains in its columns the feature maps of the sample $\{x_{t,i}\}_{i=1}^m$ used to define $\hat{m}_{ut}$. The empirical estimate of the outgoing message is then $\hat{m}_{ts}(x_s) = \left[\bigodot_{u\in\Gamma_t\setminus s}\left(K_t^{(u)}\beta_{ut}\right)\right]^\top (K_s + \lambda m I)^{-1}\Upsilon_s^\top\varphi(x_s)$. Here $\bigodot$ denotes elementwise multiplication for vectors, and we have used the empirical estimate for the conditional embedding operator and the property of tensor product features. The matrix $K_t^{(u)} := \Upsilon_t^\top\Upsilon_t^{(u)}$ is the Gram matrix between the mapped sample $\Upsilon_t^{(u)}$ used to define $\hat{m}_{ts}$ and the sample Υ_t used to define $\widehat{\mathcal{U}}_{X_t^{d_t-1}|X_s}$ (these need not be the same sample). An important point to note, when computing the message from t, is that we do *not* need to observe a joint sample across the children of t, but only samples from the *pairs* (x_u, x_t) for each $u \in \Gamma_t \setminus s$. This is essential in ensuring the resulting algorithm is useful in practice.

Algorithm 1 Inference

In: Training features Υ_s, kernel K_s, $\forall s \in \mathcal{V}$, root r.
Out: Belief B_r at root.

1: Reroot the tree at node r, orient the edges in $\mathcal{E}$
2: **for all** $t \in \mathcal{V}$ in reverse topological order **do**
3: $s = \pi(t)$
4: **if** t is the root r **then**
5: $\beta_r = \bigodot_{u \in \Gamma_r} K_r^{(u)} \beta_{ur}$
6: $B_r = \Upsilon_r \beta_r$
7: **else if** t observes evidence x_t **then**
8: $\beta_{ts} = ((K_t + \lambda I)(K_s + \lambda I))^{-1} \Upsilon_t^\top \varphi(x_t)$
9: $m_{ts} = \Upsilon_s \beta_{ts}$
10: **else if** t is an internal node **then**
11: $W_{ts} = \bigodot_{u \in \Gamma_t \setminus s} K_t^{(u)} \beta_{ut}$
12: $\beta_{ts} = (K_s + \lambda I)^{-1} W_{ts}$
13: $m_{ts} = \Upsilon_s \beta_{ts}$
14: **end if**
15: **end for**

Finally, we compute the belief at a root node r. We express this as the expectation of the mapping $\phi(x_r) \in \mathcal{F}_r$, since this gives us a conditional mean mapping, which allows us to compute expectations of any test function $g \in \mathcal{F}_r$ simply by taking an inner product. Recalling (4), we have

$$\begin{aligned} B_r &= \mathbb{E}_{X_r}\left[\varphi(X_r) \prod\nolimits_{s \in \Gamma_r} m_{sr}(X_r)\right] \\ &= \mathbb{E}_{X_r}\left[\left\langle \bigotimes\nolimits_{s \in \Gamma_r} m_{sr}, \quad \bigotimes\nolimits_{d_r} \varphi(X_r) \right\rangle_{\mathcal{H}_r} \varphi(X_r)\right], \end{aligned}$$

where we have defined $\mathcal{H}_r := \bigotimes_{d_r} \mathcal{F}_r$ as the tensor product space of d_r of the $\mathcal{F}_r$ spaces. Define $\mathcal{C}_{X_r^{d_r} X_r} := \mathbb{E}_{X_r}\left[(\bigotimes_{d_r} \varphi(X_r)) \otimes \varphi(X_r)\right]$ the covariance operator from $\mathcal{F}_r$ to $\mathcal{H}_r$, then we may express the expectation of any $g \in \mathcal{F}_r$ as

$$\langle g, B_r \rangle_{\mathcal{F}_r} = \left\langle \bigotimes\nolimits_{s \in \Gamma_r} m_{sr}, \mathcal{C}_{X_r^{d_r} X_r} g \right\rangle_{\mathcal{H}_r}.$$

Given a function $g = \Upsilon_r \beta_r$, where Υ_r contains the mapped sample used in the empirical estimate $\widehat{\mathcal{C}}_{X_r^{d_r} X_r}$, and an empirical estimate of all incoming messages $\hat{m}_{sr} = \Upsilon_r^{(s)} \beta_{sr}$, the empirical estimate $\hat{B}_r = \Upsilon_r(\bigodot_{s \in \Gamma(r)} K_r^{(s)} \beta_{sr})$ and $\left\langle g, \hat{B}_r \right\rangle_{\mathcal{F}_r} = \hat{B}_r^\top \Upsilon_r \beta_r = (\bigodot_{s \in \Gamma(r)} K_r^{(s)} \beta_{sr})^\top K_r \beta_r$, where $K_r^{(s)} = \Upsilon_r^\top \Upsilon_r^{(s)}$ and $K_r = \Upsilon_r^\top \Upsilon_r$.

One can also obtain the marginal density (or belief in the probabilistic graphical model sense) at the root node r by performing a Parzen window estimate at r. If we use a Gaussian RBF kernel, we can view the Parzen density estimator as a function in the RKHS, *i.e.*, $\mathbb{P}(x_r) = \langle \mu_r, \varphi(x_r) \rangle_{\mathcal{F}_r} = \mathbb{E}_{X_r}[k(X_r, x_r)]$ (here we have used the RKHS kernel as the Parzen window, but one could easily replace it with another kernel as long as this remains a function in $\mathcal{F}_r$). Then the belief can be estimated as $\mathbb{P}(X_r) \prod_{s \in \Gamma_r} m_{sr}(X_r) = \left\langle (\bigotimes_{s \in \Gamma_r} m_{sr}) \bigotimes \mu_r, \bigotimes_{d_r+1} \varphi(X_r) \right\rangle$. While the resulting Parzen window estimator of the belief is biased (the kernel bandwidth is fixed as a function of sample size), we will see in our experiments that we can nonetheless obtain good MAP estimators.

Algorithm: We summarize our method in Algorithm 1. The matrix inversions for $((K_t + \lambda I)(K_s + \lambda I))^{-1}$ and $(K_s + \lambda I)^{-1}$ need to be carried out only once and can be stored for future use. Suppose m is the number of training data, and n is the number of nodes in the tree. Then these once-off computation for matrix inversion requires $O(nm^3)$ floating point operations in total. The major computations of the algorithm when operating online comes from updating the weights when new pieces of evidence x_t are given (step 3, 6 and 7). Suppose the degree of node t is d_t. Then updating each β_{ts} requires $O(d_t m^2)$ operations, and therefore for all messages the algorithm requires $\sum_t O(d_t m^2) = O(nm^2)$ operations.

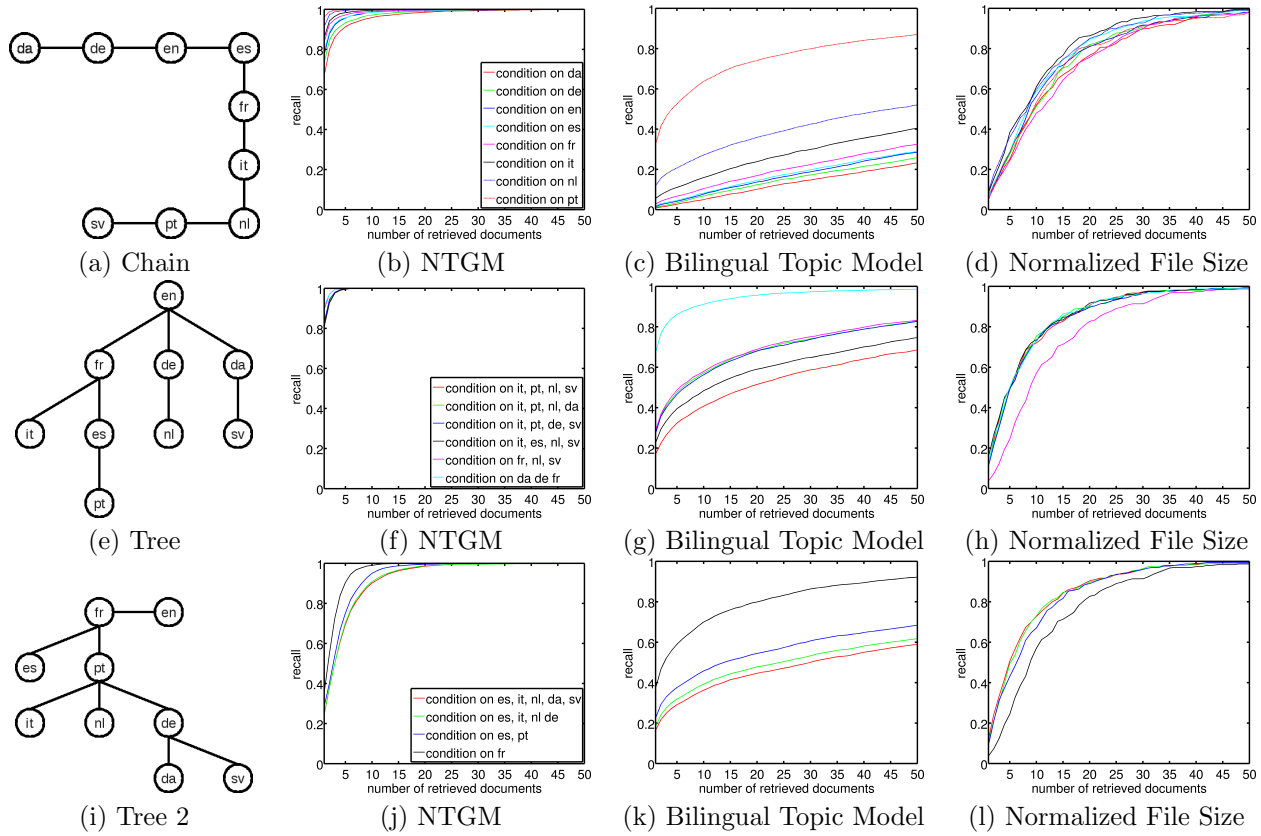
Sample complexity: We now prove that our algorithm provides convergent estimates of the beliefs at the nodes. For ease of notation, we will assume that all variables share the same domain $\mathcal{X}$, with kernel $0 \leq k(x, x') = \langle \varphi(x), \varphi(x') \rangle_{\mathcal{F}} \leq 1$. The following theorem demonstrates the convergence of the belief at the nodes s of the tree, where the proof is provided in the appendix.

Theorem 3 *For each node $s \in \mathcal{V}$, let $\mathcal{T}_s$ be the tree induced by setting node s as the root, and h_i be the depth of node i in this re-rooted tree. Let R be a constant defined in the appendix. Then subject to the assumptions required for Theorems 1 and 2 to hold for random variables on all relevant node pairs, $\frac{\|\hat{B}_s - B_s\|_{\mathcal{F}}}{\|B_s\|_{\mathcal{F}}} = O_p((\lambda^{\frac{1}{2}} + \lambda^{-2} m^{-\frac{1}{2}}) \sum_{i \in \mathcal{T}_s} R^{h_i})$.*

This bound depends on the topology of the tree: if the tree is a balanced binary tree (shallow tree), the largest exponent of R is only $\log_2(n)$ where n is $|\mathcal{V}|$; if the tree is a long chain (deep tree), the largest exponent of R will be $n - 1$. The height of the tree is related to the number of times a message is passed through an (empirically estimated) embedding operator.

6 Applications

We have applied our our nonparametric tree graphical models (NTGM) in two settings: cross-lingual document retrieval, and camera orientation recovery from images.



(a) Chain (b) NTGM (c) Bilingual Topic Model (d) Normalized File Size

(e) Tree (f) NTGM (g) Bilingual Topic Model (h) Normalized File Size

(i) Tree 2 (j) NTGM (k) Bilingual Topic Model (l) Normalized File Size

Figure 1: (a) The chain graphical model used for Swedish (sv) document retrieval. (b,c,d) The recall score for NTGM, bilingual topic model and normalized file size method for retrieving Swedish documents using languages at the front of the chain. (e) The tree graphical model used for English document retrieval. (f,g,h) The recall score for NTGM, bilingual topic model and normalized file size method for retrieval conditioned on document observations from other languages. (i) A graphical model for cross-language document retrieval, obtained via Chow-Liu with the HSIC dependence measure. (j,k,l) Recall score for NTGM, bilingual topic model and normalized file size method for retrieval conditioned on document observations from other languages.

6.1 Finding document translations

In our cross-lingual document retrieval experiments, the objective was to retrieve a document in one language given a source document in another. Our data were the proceedings of the European Parliament (Koehn, 2005). We chose 300 longest documents written in 9 languages, namely Danish (da), Dutch (nl), English (en), French (fr), German (de), Italian (it), Portuguese (pt), Spanish (es), and Swedish (sv). For this dataset, we treated each language as a random variable and each document as an observation. We constructed three graphical models over languages. Our first graph was a chain (Figure 1(a)) connecting all languages. We also produced chains of shorter length by truncating the first languages, keeping Swedish as the tail, and thus obtaining chains of length from 2 to 9. We retrieved Swedish documents given observations at the head of the chain.

Our second graph was a tree (Figure 1(e)) based on linguistic similarity of the languages, where Romance languages and Germanic languages reside in different branches. We retrieved English (en) documents conditioned on documents from other languages. We first provided evidence only on the leaf nodes (pt, it, nl and sv) and then successively added evidence to other internal nodes (es, fr, de and da). We thus retrieved each document using multiple input documents in other languages. We repeated the experiment on a third tree in Figure 1(i), obtained from the Chow-Liu algorithm using the Hilbert-Schmidt Independence Criterion (HSIC, from Gretton et al., 2005) for the required statistical dependence measure (applying the same kernels that were used in our inference algorithm). We again retrieved English documents conditioned on documents from other languages. Besides the different graph structure, all remaining experimental settings were identical to those of the linguistic similarity tree experiments in Figure 1(e). We also investigated additional tree structures, with similar results.

A key property of our method in the document retrieval application is that we do not need data for all pairs of languages, only for those connected in the graph. To learn a model on the graph, we sampled 100 matching document pairs for each pair of languages connected by an edge. The training pairs for different edges were independently sampled. Therefore, different edges might share very few common documents. We used TF-IDF (term frequency inverse document frequency) as document features, and applied a Gaussian RBF kernel, using as the bandwidth the median distance between the feature vectors. For preprocessing we removed stopwords (http://www.nltk.org) and performed stemming (http://snowball.tartarus.org). We retrieved up to 50 documents, and evaluated the performance using the recall score. All 300 documents were used as test queries. We randomized the experiments 10 times and report the average score.

We compared our approach against two baselines. The first was to use document length directly as a sole feature for retrieval, since length will be roughly retained across languages (see also the sentence alignment work of Gale & Church, 1991). Since some languages are less terse than others, we normalized the within-language document length to zero mean and unit variance.

As our second point of comparison, we employed a polylingual topic model for document tuples (Mimno et al., 2009). We learned a bilingual topic model for each edge based on the training document pairs (with stemming and stopword removal). We trained each bilingual topic model with 50 topics using Gibbs sampling, and inferred the topic distribution for each query document. Document retrieval in the target domain was achieved by comparing the topic distributions of the query and target documents. We used the Jensen-Shannon divergence for this comparison. Since bilingual topic models were built only for each edge, we needed to perform multiple intermediate retrievals.

Experimental results are summarized in Figure 1. For the chain, the performance of NTGM and bilingual topic models degrades with increased length. The performance of bilingual topic models drops significantly faster than our method, however. For the retrieval method based on file size, the result is surprisingly good, and the performance is not much affected by the length of the chain. However, the file size based method does not perform better than our method when the chain is short.

For the tree in Figure 1(e), our method achieves very good results. As the evidence moves closer to the root (English), retrieval performance improves. Bilingual LDA produces a qualitatively similar result, but its performance is much worse than NTGM. Again, retrieval using file size performs well. That said, when the evidence is moved further towards the root, performance remains stagnant. Results for the third tree obtained via Chow-Liu are provided in Figures 1(j,k,l), and are qualitatively similar to the cross-language retrieval results using the linguistic similarity tree.

6.2 Finding camera rotations

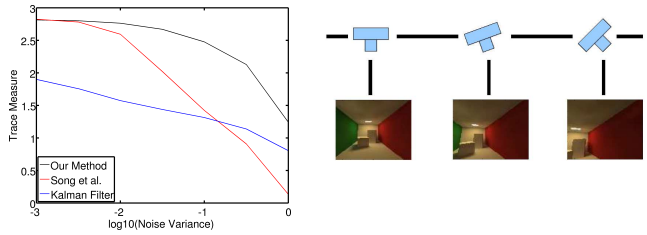
We applied NTGM to to a computer vision problem as in Song et al. (2009). Our goal was to determine the camera orientation based on the images it observed. In this setting, the camera focal point was fixed at a position and traced out a smooth path of rotations while making observations. The dataset was generated by POVRAY,[5] which rendered images observed by the camera. The virtual scene was a rectangular-shaped room with a ceiling light and two pieces of furniture. The images exhibited complex lighting effects such as shadows, interreflections, and global illumination, all of which made determining the camera rotation difficult, especially for noisy cases.

The sequence of image observations contained 3600 frames, where we used the first 1800 frames for training and the remaining 1800 frames for testing. The dynamics governing the camera rotation were characterized by a piece-wise smooth random walk. This is an unconventional graphical model in that the camera state is a rotation matrix R from $SO(3)$; and the observations are images which are high dimensional spaces with correlation between pixel values. The graph structure for this problem is the caterpillar tree shown in Figure 2(b), and we perform online inference.

We flattened each image to a vector, and applied a Gaussian RBF kernel. The bandwidth parameter of the kernel was fixed to be the median distance between image vectors. We used a Gaussian RBF kernel between two rotations R and $\tilde{R}$, *i.e.*, $k(R, \tilde{R}) := \exp(-\sigma\|R - \tilde{R}\|^2)$. With this kernel, we found the most probable camera rotation matrix by maximizing the belief $B(R)$ over the rotation group (Abrudan et al., 2008).

We compared our method to a Kalman filter, and to the method of Song et al. (2009). For the Kalman filter, we used the quaternion corresponding to a rotation matrix R as the state and the image vectors as the observations. We learned the model parameters of the linear dynamical system using linear regression. In Song et al., a simplifying approximation was made in aggregating dynamical system history and the current image observation. We expect NTGM, which incorporates both sources of information in a princi-

[5] www.povray.org



(a) (b)

Figure 2: Performance of different methods vs observation noise, camera rotation problem.

pled way, to outperform this earlier method. We used $\mathrm{tr}(R^\top \hat{R})$ between the true rotation R and the estimated one $\hat{R}$ as performance measure (this measure ranges between $[-1, 3]$: larger means better performance).

We added zero mean Gaussian white noise to the images and measured the performance scaling of the three methods as we increased the noise variance. The performance of NTGM degrades more slowly than the other two methods (Figure 2(a)). For large noise, the Kalman filter overtakes the method of Song et al. In this setting, the images are very noisy, and the dynamics become key to determining the camera orientation. In this regime, NTGM significantly outperforms the other two methods, with 40% higher trace measure.

7 Conclusion

We propose a nonparametric representation for graphical models on trees. The key to this new framework is to represent marginals, conditionals, and likelihoods as Hilbert space elements. Based on these representations, we can manipulate the messages and beliefs of the model solely on the basis of linear operations in feature space, and we derive an efficient kernel belief propagation algorithm for performing inference. Our approach can handle challenging problems with continuous and structured variables, which are difficult for alternative methods. In our experiments, we applied this nonparametric tree model to a document retrieval task and a camera rotation estimation problem. In both cases, our algorithm outperforms state-of-the-art techniques. We anticipate the application of kernel methods to inference in graphical models will be fruitful not only in broadening the classes of data on which inference is tractable, but also in generalizing kernel techniques to more complex dependence structures.

Acknowledgements We thank Kenji Fukumizu and Alex Smola for helpful discussions. This work was supported by grants DARPA IPTO FA8750-09-1-0141, ONR MURI N000140710747 and NSF NeTS-NOSS CNS-0625518.

References

Abrudan, T., Eriksson, J., & Koivunen, V. (2008). Steepest descent algorithms for optimization under unitary matrix constraint. *IEEE SP*, *56*(3).

Baker, C. (1973). Joint measures and cross-covariance operators. *Trans. AMS*, *186*, 273–289.

Berlinet, A., & Thomas-Agnan, C. (2003). *Reproducing Kernel Hilbert Spaces in Probability and Statistics*. Berlin: Springer-Verlag.

Fukumizu, K., Bach, F., & Gretton, A. (2007). Statistical consistency of kernel canonical correlation analysis. *JMLR*, *8*, 361–383.

Fukumizu, K., Bach, F. R., & Jordan, M. I. (2004). Dimensionality reduction for supervised learning with RKHSs. *JMLR*, *5*, 73–99.

Gale, W. A., & Church, K. W. (1991). A program for aligning sentences in bilingual corpora. In *Meeting of the Association for Computational Linguistics*, 177–184.

Gretton, A., Borgwardt, K., Rasch, M., Schölkopf, B., & Smola, A. (2007). A kernel method for the two-sample-problem. In *NIPS*.

Gretton, A., Bousquet, O., Smola, A., & Schölkopf, B. (2005). Measuring statistical dependence with Hilbert-Schmidt norms. In *ALT 16*, 63–78.

Ihler, A., & McAllester, D. (2009). Particle belief propagation. In *AISTATS*.

Jordan, M. I. (2002). *An Introduction to Probabilistic Graphical Models*. MIT Press. To Appear.

Koehn, P. (2005). Europarl: A parallel corpus for statistical machine translation. In *MTS X*, 79–86.

Mimno, D., Wallach, H., Naradowsky, J., Smith, D., & McCallum, A. (2009). Polylingual topic models. In *EMNLP*.

Minka, T. (2001). Expectation Propagation for Approximate Bayesian Inference. In *UAI*.

Pearl, J. (1988). *Probabilistic Reasoning in Intelligent Systems*. Morgan-Kaufman.

Schölkopf, B., Tsuda, K., & Vert, J.-P. (2004). *Kernel Methods in Computational Biology*. MIT Press.

Smola, A., Gretton, A., Song, L., & Schölkopf, B. (2007). A Hilbert space embedding for distributions. In *ALT*.

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

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

Sudderth, E., Ihler, A., Freeman, W., & Willsky, A. (2003). Nonparametric belief propagation. In *CVPR*.

Taskar, B., Guestrin, C., & Koller, D. (2004). Max-margin Markov networks. In *NIPS*.

Weiss, Y., & Freeman, W. (2001). Correctness of belief propagation in gaussian graphical models of arbitrary topology. *Neural Computation*, *13*(10), 2173–2200.

Supplementary to Nonparametric Tree Graphical Models via Kernel Embeddings

Le Song, Arthur Gretton, Carlos Guestrin

April 1, 2010

The appendix contains proofs of the main theorems.

1 Preliminary results

Given any operator $A : \mathcal{G} \to \mathcal{F}$, the operator norm of A is written $\|A\|_2$, and its Hilbert-Schmidt norm (where defined) is

$$\|A\|_{HS}^2 := \sum_{i,j=1}^{\infty} \langle \varphi_j, A\phi_i \rangle_{\mathcal{F}}^2 ,$$

where the φ_i form a complete orthonormal system (CONS) for $\mathcal{F}$, and the ϕ_j form a CONS for $\mathcal{G}$. The set of Hilbert-Schmidt operators has the inner product

$$\langle A, B \rangle_{HS} = \sum_{i,j \geq 1} \langle A\phi_i, \varphi_j \rangle_{\mathcal{F}} \langle B\phi_i, \varphi_j \rangle_{\mathcal{F}}$$

We have defined the rank one operator $f \otimes g : \mathcal{G} \to \mathcal{F}$ such that $f \otimes g(h) = \langle g, h \rangle_{\mathcal{G}} f$.

It follows that

$$\langle f \otimes g, A \rangle_{HS} = \langle Ag, f \rangle_{\mathcal{F}} ,$$

and in particular,

$$\langle a \otimes b, u \otimes v \rangle_{HS} = \langle a, u \rangle_{\mathcal{F}} \langle b, v \rangle_{\mathcal{G}} .$$

We can extend this notation to higher order: for instance, given the product space $\mathcal{F}^n$ and functions $a_i \in \mathcal{F}$ and $b_i \in \mathcal{F}$ for $i \in \{1, \ldots, n\}$,

$$\left\langle \bigotimes_{i=1}^{n} a_i, \bigotimes_{i=1}^{n} b_i \right\rangle_{\mathcal{F}^n} = \prod_{i=1}^{n} \langle a_i, b_i \rangle_{\mathcal{F}} . \tag{1}$$

We use the result

$$A^{-1} - B^{-1} = A^{-1}(B - A)B^{-1}. \tag{2}$$

Further, following [2], we may define the empirical regularized correlation operator $\hat{\mathcal{V}}_{XY}$ such that

$$\hat{\mathcal{C}}_{XY} := \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{1/2} \hat{\mathcal{V}}_{XY} \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{1/2}. \tag{3}$$

where we have $\left\|\hat{\mathcal{V}}_{XY}\right\| \leq 1$.

2 Proof of Theorem 1

We now prove the result

$$\left\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\right\|_{HS} = O_p(\lambda_m^{\frac{1}{2}} + \lambda_m^{-\frac{3}{2}} m^{-\frac{1}{2}}). \tag{4}$$

We define a regularized population operator

$$\tilde{\mathcal{U}}_{Y|X} := \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}$$

and decompose (4) as

$$\left\|\hat{\mathcal{U}}_{Y|X} - \mathcal{U}_{Y|X}\right\|_{HS} \leq \left\|\hat{\mathcal{U}}_{Y|X} - \tilde{\mathcal{U}}_{Y|X}\right\|_{HS} + \left\|\mathcal{U}_{Y|X} - \tilde{\mathcal{U}}_{Y|X}\right\|_{HS}.$$

There are two parts to the proof. In the first part, we show convergence in probability of the first term in the above sum. In the second part, we demonstrate that as long as $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ is Hilbert-Schmidt, the second term in the sum converges to zero as λ_m drops.

Part 1: We make the decomposition

$$\begin{aligned} & \left\|\mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1} - \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1}\right\|_{HS} \\ \leq & \left\|\left(\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right) \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}\right\|_{HS} + \left\|\hat{\mathcal{C}}_{YX} \left[\left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1} - \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}\right]\right\|_{HS}. \end{aligned}$$

The first term is bounded according to

$$\left\|\left(\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right) \left(\mathcal{C}_{XX} + \lambda_m I\right)^{-1}\right\|_{HS} \leq \frac{1}{\lambda_m} \left\|\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right\|_{HS},$$

and we know from [1, Lemma 5] that $\left\|\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}\right\|_{HS} = O_p(1/\sqrt{m})$. For the

second term, we first substitute (2) and then (3) to obtain

$$\begin{aligned}
& \left\| \hat{\mathcal{C}}_{YX} \left[\left(\hat{\mathcal{C}}_{XX} + \lambda_m I \right)^{-1} - \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right] \right\|_{HS} \\
= & \left\| \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{XX} + \lambda_m I \right)^{-1} \left[\mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right] \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS} \\
= & \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I \right)^{1/2} \hat{\mathcal{V}}_{XY} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1/2} \left[\mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right] \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS} \\
\leq & \frac{\left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I \right)^{1/2} \right\|}{\lambda_m^{3/2}} \left\| \mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right\|_{HS} = O_p(\lambda_m^{-\frac{3}{2}} m^{-\frac{1}{2}}).
\end{aligned}$$

Part 2: $\left\| \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} - \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS} = O(\lambda_m^{\frac{1}{2}})$.

Proof: We first expand the covariance operator $\mathcal{C}_{XX}$ in terms of the complete orthonormal system (CONS)

$$\mathcal{C}_{XX} = \sum_{i=1}^{\infty} \nu_i \varphi_i \otimes \varphi_i. \tag{5}$$

Then

$$\begin{aligned}
& \left\| \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} - \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right\|_{HS}^2 \qquad (6) \\
= & \sum_{i,j=1}^{\infty} \left\langle \phi_j, \left(\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1} - \mathcal{C}_{YX} \left(\mathcal{C}_{XX} + \lambda_m I \right)^{-1} \right) \varphi_i \right\rangle^2 \\
= & \sum_{i,j=1}^{\infty} \left\langle \phi_j, \mathcal{C}_{YX}\nu_i^{-1}\varphi_i - \mathcal{C}_{YX}(\lambda_m + \nu_i)^{-1}\varphi_i \right\rangle^2 \\
= & \sum_{i,j=1}^{\infty} \left(\frac{\lambda_m}{\nu_i + \lambda_m} \right)^2 \left\langle \phi_j, \mathcal{C}_{YX}\nu_i^{-1}\varphi_i \right\rangle^2 \\
= & \sum_{i,j=1}^{\infty} \left(\frac{\lambda_m}{\nu_i + \lambda_m} \right)^2 \left\langle \phi_j, \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\varphi_i \right\rangle^2
\end{aligned}$$

Next, define

$$s_{ji} := \langle \phi_j, \mathcal{C}_{YX}\varphi_i \rangle$$

Assuming $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$ is Hilbert-Schmidt, we have that

$$\sum_{i,j=1}^{\infty} \left\langle \phi_j, \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}\varphi_i \right\rangle^2 = \sum_{i,j=1}^{\infty} \frac{s_{ji}^2}{\nu_i^2} \quad \text{is finite.}$$

Furthermore,

$$\left(\frac{\lambda_m}{\nu_i + \lambda_m} \right)^2 = \left(\frac{1}{\frac{1}{\frac{\lambda_m}{\nu_i}} + \frac{1}{1}} \right)^2 \leq \left(\frac{1}{2} \sqrt{\frac{\lambda_m}{\nu_i}} \right)^2 = \frac{1}{4} \frac{\lambda_m}{\nu_i}$$

where we have used the arithmetic-geometric-harmonic means inequality. Therefore we need

$$\sum_{i,j=1}^{\infty} \frac{1}{4} \frac{\lambda_m}{\nu_i} \frac{s_{ji}^2}{\nu_i^2} \quad \text{to be finite.}$$

If we assume that

$$c := \sum_{i,j=1}^{\infty} \frac{1}{4} \frac{s_{ji}^2}{\nu_i^3} \quad \text{is finite,}$$

which corresponds to $\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ being Hilbert-Schmidt, then the squared norm difference in (6) will approach zero with rate $\lambda_m c$.

3 Proof of Theorem 2

We make a similar decomposition to the proof of Theorem 1, yielding

$$\begin{aligned}
& \left\| (\mathcal{C}_{YY} + \lambda_m I)^{-1} \mathcal{C}_{YX} (\mathcal{C}_{XX} + \lambda_m I)^{-1} - \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \hat{\mathcal{C}}_{YX} \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1} \right\|_{HS} \\
\leq & \left\| \left[(\mathcal{C}_{YY} + \lambda_m I)^{-1} - \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \right] \mathcal{C}_{YX} (\mathcal{C}_{XX} + \lambda_m I)^{-1} \right\|_{HS} \\
& + \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} (\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}) (\mathcal{C}_{XX} + \lambda_m I)^{-1} \right\|_{HS} \\
& + \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \hat{\mathcal{C}}_{YX} \left[(\mathcal{C}_{XX} + \lambda_m I)^{-1} - \left(\hat{\mathcal{C}}_{XX} + \lambda_m I\right)^{-1} \right] \right\|_{HS}.
\end{aligned}$$

The first term is bounded according to

$$\begin{aligned}
& \left\| \left[(\mathcal{C}_{YY} + \lambda_m I)^{-1} - \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \right] \mathcal{C}_{YX} (\mathcal{C}_{XX} + \lambda_m I)^{-1} \right\|_{HS} \\
= & \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \left[\mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY} \right] (\mathcal{C}_{YY} + \lambda_m I)^{-1} \mathcal{C}_{YX} (\mathcal{C}_{XX} + \lambda_m I)^{-1} \right\|_{HS} \\
\leq & \left\| \left(\hat{\mathcal{C}}_{YY} + \lambda_m I\right)^{-1} \left[\mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY} \right] (\mathcal{C}_{YY} + \lambda_m I)^{-1/2} V_{XY} (\mathcal{C}_{XX} + \lambda_m I)^{-1/2} \right\|_{HS} \\
\leq & \frac{\left\| \mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY} \right\|_{HS}}{\lambda_m^2} = O_p(\lambda_m^{-2} m^{-\frac{1}{2}}).
\end{aligned}$$

The third term follows similar reasoning. The second term is bounded according to

$$\left\| (\hat{\mathcal{C}}_{YY} + \lambda_m I)^{-1} (\mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX}) (\mathcal{C}_{XX} + \lambda_m I)^{-1} \right\|_{HS} \leq \frac{\left\| \mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX} \right\|_{HS}}{\lambda_m^2} = O_p(\lambda_m^{-2} m^{-\frac{1}{2}}).$$

Convergence in probability of the three terms follows from the convergence of each of $\left\| \mathcal{C}_{YY} - \hat{\mathcal{C}}_{YY} \right\|_{HS}$, $\left\| \mathcal{C}_{YX} - \hat{\mathcal{C}}_{YX} \right\|_{HS}$, and $\left\| \mathcal{C}_{XX} - \hat{\mathcal{C}}_{XX} \right\|_{HS}$, as in the proof of Theorem 1.

We next address the convergence of

$$\left\|\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}\left(\mathcal{C}_{XX}+\lambda_m I\right)^{-1}-\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}\right\|_{HS}.$$

for λ_m approaching zero. We use the earlier decomposition of $\mathcal{C}_{XX}$ in terms of its eigenfunctions φ_i from (5), and further require that ϕ_i be the eigenfunctions of $\mathcal{C}_{YY}$,

$$\mathcal{C}_{YY}:=\sum_{i=1}^{\infty}\gamma_i\phi_i\otimes\phi_i.$$

Thus

$$\begin{aligned}
&\left\|\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}-\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}\left(\mathcal{C}_{XX}+\lambda_m I\right)^{-1}\right\|_{HS}^2\\
=&\sum_{i,j=1}^{\infty}\left\langle\phi_j,\left(\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}-\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}\left(\mathcal{C}_{XX}+\lambda_m I\right)^{-1}\right)\varphi_i\right\rangle^2\\
=&\sum_{i,j=1}^{\infty}\left\langle\phi_j,\mathcal{C}_{YY}^{-1}C_{XY}\nu_i^{-1}\varphi_i-\left(\mathcal{C}_{YY}+\lambda_m I\right)^{-1}\mathcal{C}_{YX}(\nu_i+\lambda_m)^{-1}\varphi_i\right\rangle^2\\
=&\sum_{i,j=1}^{\infty}\left\langle\phi_j,C_{XY}\left(\gamma_j\nu_i\right)^{-1}\varphi_i-\mathcal{C}_{YX}(\nu_i+\lambda_m)^{-1}(\gamma_j+\lambda_m)^{-1}\varphi_i\right\rangle^2\\
=&\sum_{i,j=1}^{\infty}\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\left(\nu_i+\lambda_m\right)\left(\gamma_j+\lambda_m\right)}\right)^2\left\langle\phi_j,\mathcal{C}_{YY}^{-1}C_{XY}\mathcal{C}_{XX}^{-1}\varphi_i\right\rangle^2.
\end{aligned}$$

Furthermore, we have

$$\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\nu_i\gamma_j+\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}\right)^2\leq\frac{1}{4}\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\nu_i\gamma_j}\right),$$

where we again use the arithmetic-geometric-harmonic mean inequality. Assuming $\lambda_m\ll\gamma_1$ and $\lambda_m\ll\nu_1$, it follows that

$$\lambda_m^2<\gamma_1\lambda_m+\nu_1\lambda_m,$$

and thus

$$\frac{1}{4}\left(\frac{\lambda_m^2+\gamma_j\lambda_m+\nu_i\lambda_m}{\nu_i\gamma_j}\right)<\frac{1}{2}\left(\frac{\gamma_1\lambda_m+\nu_1\lambda_m}{\nu_i\gamma_j}\right).$$

We therefore require the finiteness of

$$\sum_{i,j=1}^{\infty}\frac{\lambda_m}{2}\left(\frac{\nu_1+\gamma_1}{\nu_i\gamma_j}\right)\frac{s_{ij}^2}{\nu_i^2\gamma_j^2}<\frac{\lambda_m\left(\nu_1+\gamma_1\right)}{2}\sum_{i,j=1}^{\infty}\frac{s_{ij}^2}{\nu_i^3\gamma_j^3}.$$

This is equivalent to requiring that $\mathcal{C}_{YY}^{-\frac{3}{2}}\mathcal{C}_{YX}\mathcal{C}_{XX}^{-\frac{3}{2}}$ be Hilbert-Schmidt as a condition of convergence.

4 Proof of Theorem 3

Our bound is in terms of the following constants:

$$R_m = \max_{(s,t)\in\mathcal{E}} \frac{\|M_{ts}\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \tag{7}$$

$$R_B = \max_{t\in\mathcal{V}} \frac{\|M_{ts}\otimes m_{st}\|_{\mathcal{H}_t\otimes\mathcal{F}}}{\|B_t\|_{\mathcal{F}}} \tag{8}$$

$$R_{\mathcal{L}} = \max_{(s,t)\in\mathcal{E}} \sup_{x_t\in\mathcal{X}} \|f_{x_t}\|_{\mathcal{F}}^{-1} \tag{9}$$

$$R = \max\{R_m, R_B, R_{\mathcal{L}}\} \tag{10}$$

where R_m is the maximal ratio of the RKHS norm of the pre-message to that of the message, R_B is maximal ratio of the RKHS norm of the pre-belief to that of the belief, and $R_{\mathcal{L}}$ is the maximal inverse of the RKHS norm of f_{x_t}. R is the largest of these three quantities. R_m and R_B quantify the degree of smoothing of the RKHS function after message propagation, while $R_{\mathcal{L}}$ quantifies the smoothness of the RKHS function f_{x_t} itself. Under our assumption that $0 \leq k(x,x') = \langle\varphi(x),\varphi(x')\rangle_{\mathcal{F}} \leq 1$, we have $\|\hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\|_2 \leq 1$ and $\left\|\mathcal{C}_{X_s^{d_s}X_s}\right\|_2 \leq 1$.

Proof We first bound the difference between the true message $m_{ts} = M_{ts}^\top \mathcal{U}_{X_t|X_s}$ and the message produced by propagating the true "pre-message" through the estimated embedding operator $\tilde{m}_{ts} := M_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}$:

$$\begin{aligned}
\frac{\|\tilde{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} &= \frac{\left\|M_{ts}^\top \mathcal{U}_{X_t^{d_t-1}|X_s} - M_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\right\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} \\
&\leq \frac{\|M_{ts}\|_{\mathcal{H}_t}}{\|m_{ts}\|} \left\|\mathcal{U}_{X_t^{d_t-1}|X_s} - \hat{\mathcal{U}}_{X_t^{d_t-1}X_s}\right\|_{HS} \\
&\leq RC\left(\tfrac{\delta}{2(n-1)}\right)\lambda^{-2}m^{-\frac{1}{2}} =: \epsilon
\end{aligned} \tag{11}$$

with probability at least $1-\delta$ simultaneously for all $2(n-1)$ messages, using the union bound. The first inequality follows from $\|\mathcal{T}a\|_{\mathcal{F}} \leq \|\mathcal{T}\|_2 \|a\|_{\mathcal{F}}$, and the relation between the spectral norm and Hilbert-Schmidt norm of operators, *i.e.* $\|\mathcal{T}\|_2 \leq \|\mathcal{T}\|_{HS}$. We then have

$$\tilde{m}_{ts} \in m_{ts} + v\cdot\epsilon\|m_{ts}\|_{\mathcal{F}}, \quad \|v\|_{\mathcal{F}} \leq 1 \tag{12}$$

Note that $\tilde{m}_{ts}$ is different from the estimated message $\hat{m}_{ts}(x_s) := \hat{M}_{ts}^\top \hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\varphi(x_s)$, where both the pre-message and the conditional embedding operator are esti-

mated. Next, we bound

$$\begin{aligned}\frac{\|\hat{m}_{ts}-m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} &\leq \frac{\|\hat{m}_{ts}-\tilde{m}_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}}+\frac{\|\tilde{m}_{ts}-m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} \\ &\leq \frac{\left\|\hat{M}_{ts}^{\top}\hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}-M_{ts}^{\top}\hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\right\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}}+\epsilon \\ &\leq \frac{\left\|\hat{M}_{ts}-M_{ts}\right\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}}+\epsilon \end{aligned} \tag{13}$$

where we use $\|\hat{\mathcal{U}}_{X_t^{d_t-1}|X_s}\|_2 \leq 1$. Furthermore, we have:

$$\begin{aligned}&\frac{\left\|\hat{M}_{ts}-M_{ts}\right\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \\ =&\frac{\|\bigotimes_u(m_{ut}+v_u\cdot\epsilon_u\|m_{ut}\|)-\bigotimes_u m_{ut}\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}} \\ =&\frac{\|M_{ts}\|_{\mathcal{H}_t}}{\|m_{ts}\|_{\mathcal{F}}}\left\|\bigotimes\nolimits_u(w_u+v_u\cdot\epsilon_u)-\bigotimes\nolimits_u w_u\right\|_{\mathcal{H}_t} \\ &\left(\|M_{ts}\|_{\mathcal{H}_t}=\prod\nolimits_u\|m_{ut}\|_{\mathcal{F}} \text{ and } \|w_u\|_{\mathcal{F}}=1\right) \\ \leq& R\left\|\bigotimes\nolimits_u w_u(1+\epsilon_u)-\bigotimes\nolimits_u w_u\right\|_{\mathcal{H}_t} \\ \leq& R\left(\prod\nolimits_u(1+\epsilon_u)-1\right) \\ =& R\left(\sum_u\epsilon_u+\sum_{u,u'}O(\epsilon_u\epsilon_{u'})\right)\end{aligned} \tag{14}$$

We can then prove by induction that

$$\frac{\|\hat{m}_{ts}-m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} \leq \epsilon\sum_{i\in\mathcal{T}_t}R^{h_i}+O(\epsilon^2) =: \epsilon_t \tag{15}$$

where $\mathcal{T}_t$ is the subtree induced by node t when it sends a message to s. For a node i in the subtree $\mathcal{T}_t$, h_i denotes the depth of this node. The root node of the subtree $\mathcal{T}_t$, *i.e.* node t, starts with depth 0, *i.e.* $h_t = 0$.

For a leaf node, the subtree $\mathcal{T}_t$ contains a single node, and $m_{ts} = f_{x_t}$. We have

$$\frac{\left\|\hat{f}_{x_t}-f_{x_t}\right\|_{\mathcal{F}}}{\|f_{x_t}\|_{\mathcal{F}}} \leq \frac{\left\|\hat{\mathcal{A}}_{ts}-\mathcal{A}_{ts}\right\|_{HS}}{\|f_{x_t}\|_{\mathcal{F}}} \leq \epsilon. \tag{16}$$

Assume that (15) holds for all messages coming into node t. Combining (13)

and (14),

$$\begin{aligned}\frac{\|\hat{m}_{ts} - m_{ts}\|_{\mathcal{F}}}{\|m_{ts}\|_{\mathcal{F}}} &\leq \epsilon \sum_{u} \sum_{i \in \mathcal{T}_u} R^{h_i+1} + O(\epsilon^2) \\ &= \epsilon \sum_{j \in \mathcal{T}_t} R^{h_j} + O(\epsilon^2) \end{aligned} \tag{17}$$

where in the last equality we have grown the tree by one level. Applying a similar argument to the final belief B_s and using $\left\|\mathcal{C}_{X_s^{d_s} X_s}\right\|_2 \leq 1$, we complete the proof. ■

References

[1] K. Fukumizu, F. Bach, and A. Gretton. Statistical consistency of kernel canonical correlation analysis. *JMLR*, 8:361–383, 2007.

[2] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf. Kernel measures of conditional dependence. In *Advances in Neural Information Processing Systems 20*, pages 489–496, Cambridge, MA, 2008. MIT Press.