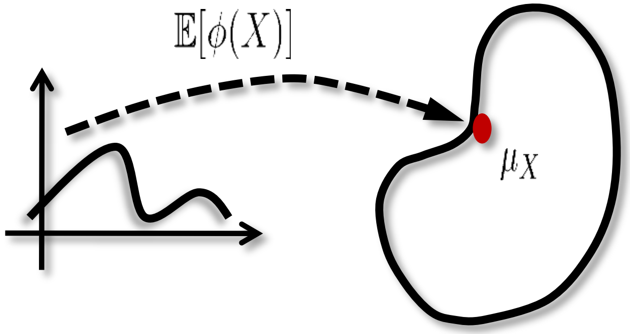
School of Computer Science
Carnegie Mellon

Probabilistic Graphical Models

Introduction to Hilbert Space Embeddings

Ankur Parikh

Lecture 19, March 27, 2013



The Optimization View of Graphical Models

- The connection between optimization and graphical models has led to many amazing discoveries
  - EM
  - Variational Inference
  - Max Margin/Max Entropy Learning
  - Bridge to Statistical Physics, Numerical Methods Communities

- Optimization has many advantages:
  - It is easy to formulate
  - Can derive principled approximations via convex relaxations
  - Can use existing optimization methods.

- But it has many challenges too:
  - Non-Gaussian continuous variables
  - Nonconvexity (local minima)

The Linear Algebra View of Graphical Models

- We are going to discuss a different (still not fully understood) point of view of graphical models that revolves around linear algebra.

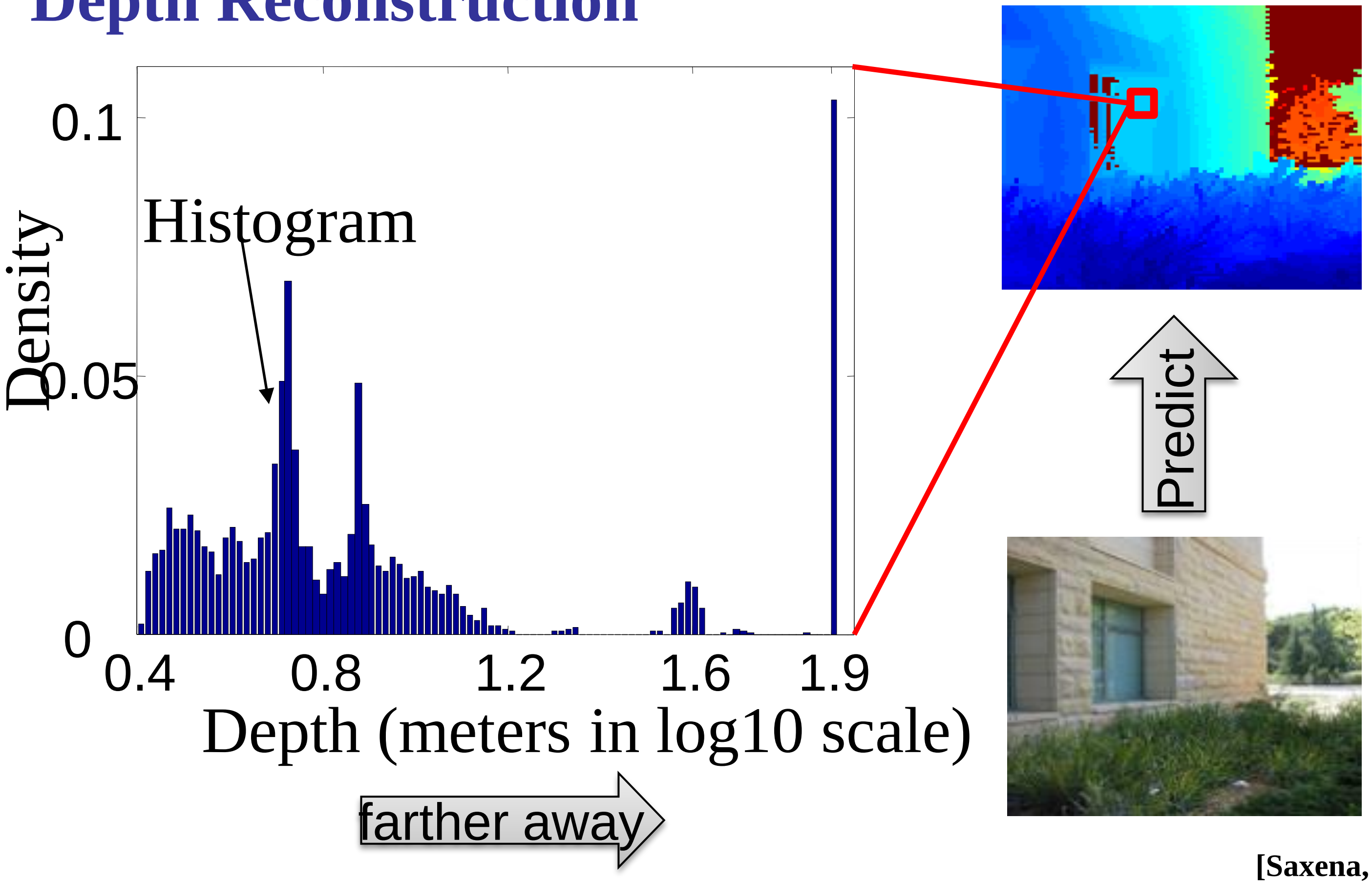
- Compared to the optimization perspective, the linear algebra view often less intuitive to formulate.

- However, it lets us solve problems that are intractable from the optimization perspective
    - Graphical Models with Non-Gaussian Continuous Variables.
    - Local Minima Free Learning in Latent Variable Models

- Moreover it offers a different theoretical perspective and bridges the graphical models, kernels and tensor algebra communities.

Non-Gaussian Continuous Variables

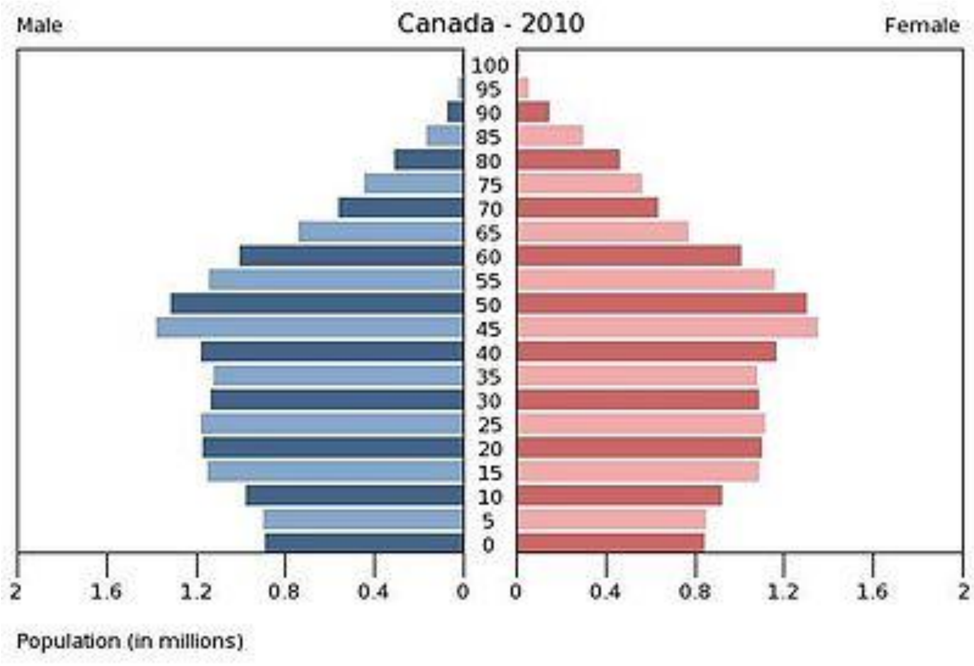
Depth Reconstruction

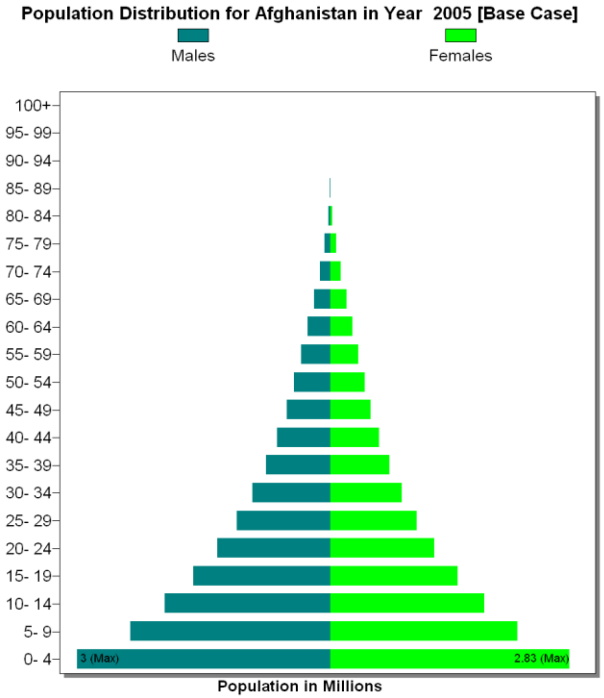


[Saxena, Chung and Ng 2005]

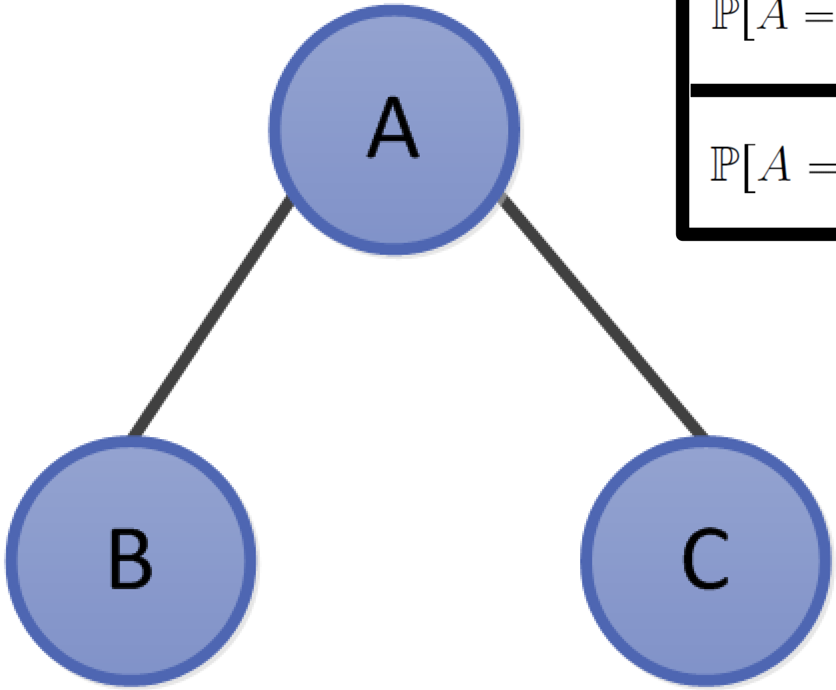
Non-Gaussian Continuous Variables

Demographics: Model relationships among different demographic variables





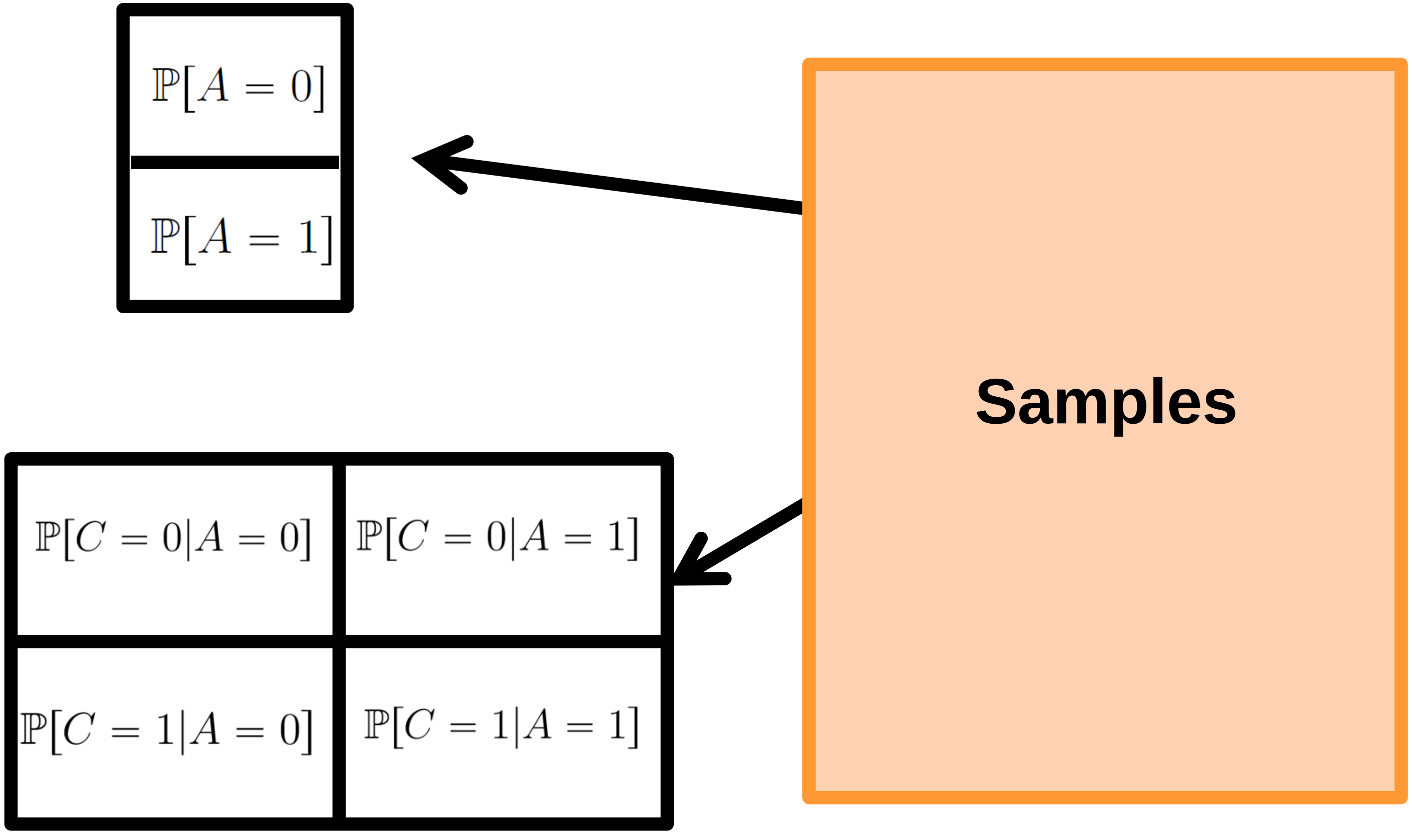
Graphical Models - What we have learned so far…



$\mathbb{P}[A=0]$
$\mathbb{P}[A=1]$

$\mathbb{P}[C=0 \mid A=0]$	$\mathbb{P}[C=0 \mid A=1]$
$\mathbb{P}[C=1 \mid A=0]$	$\mathbb{P}[C=1 \mid A=1]$

Parameter Learning - What we have learned so far…



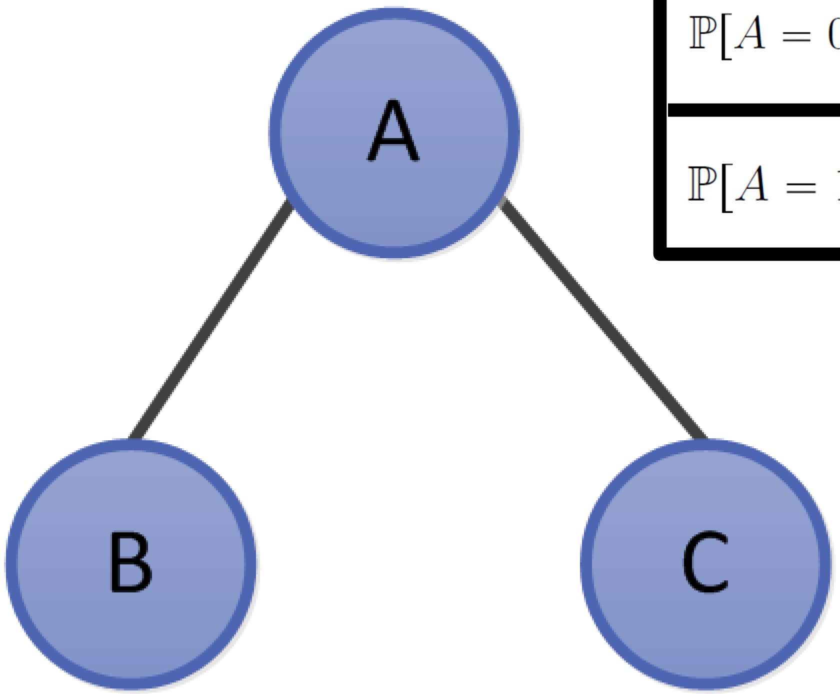
$\mathbb{P}[A = 0]$
$\mathbb{P}[A = 1]$

$\mathbb{P}[C = 0 \vert A = 0]$	$\mathbb{P}[C = 0 \vert A = 1]$
$\mathbb{P}[C = 1 \vert A = 0]$	$\mathbb{P}[C = 1 \vert A = 1]$

Samples

- If variables are observed, just count from dataset
- In case of hidden variables, can use Expectation Maximization…..

Inference - What we have learned so far…



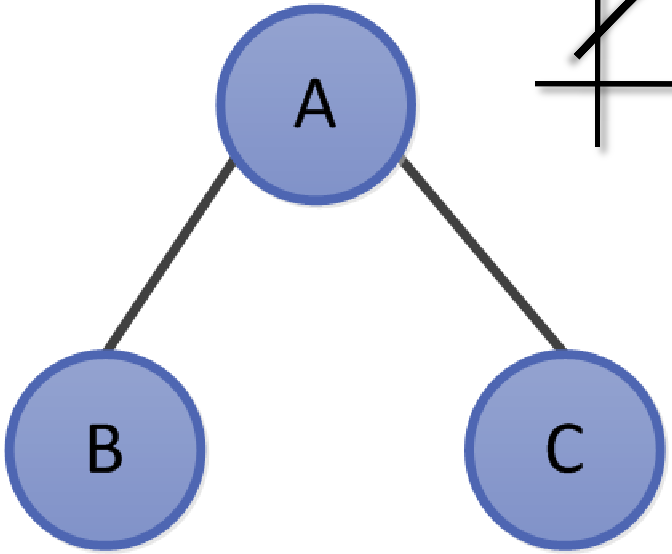
$\mathbb{P}[A=0]$
$\mathbb{P}[A=1]$

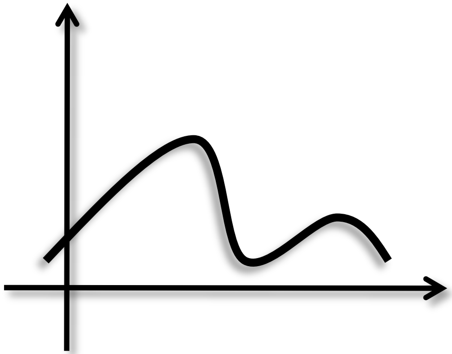
$\mathbb{P}[C=0\|A=0]$	$\mathbb{P}[C=0\|A=1]$
$\mathbb{P}[C=1\|A=0]$	$\mathbb{P}[C=1\|A=1]$

- Can do exact inference with Variable Elimination, Belief Propagation.

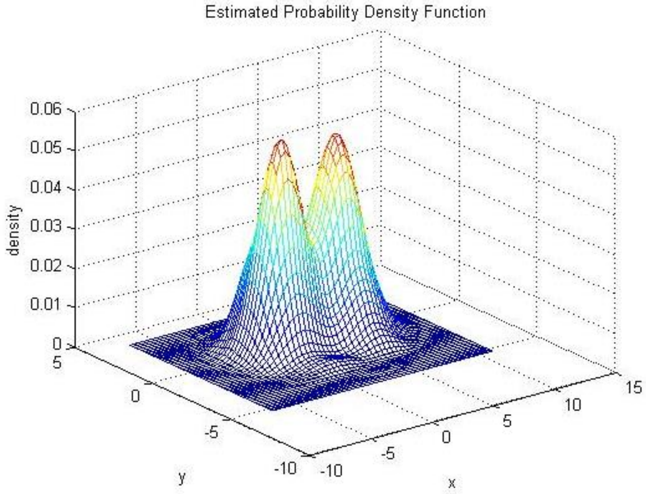
- Can do approximate inference with Loopy BP, Mean Field, MCMC

Non-Parametric Continuous Case is Much Harder…



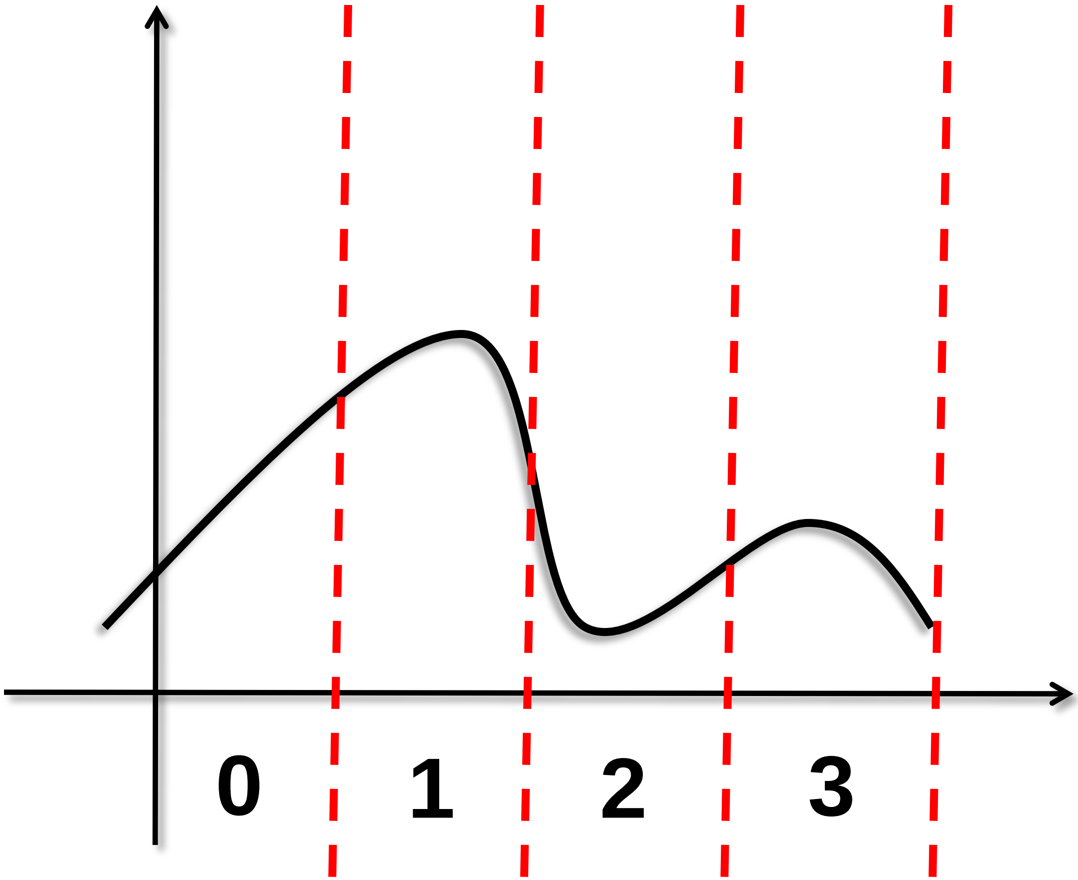


How do we make a conditional probability table out of this?



- **How to learn parameters? (What are the parameters?)**
- **How to perform inference?**

Could Discretize the Distribution….



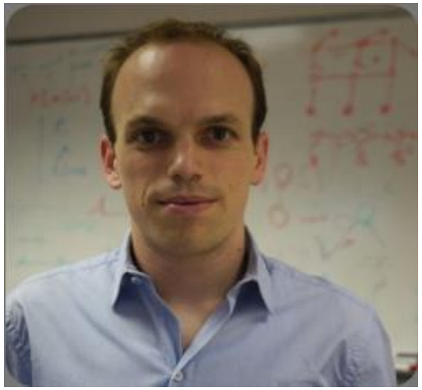
- **Loses information** that 0 and 1 are closer than 0 and 3

Hilbert Space Embeddings of Distributions

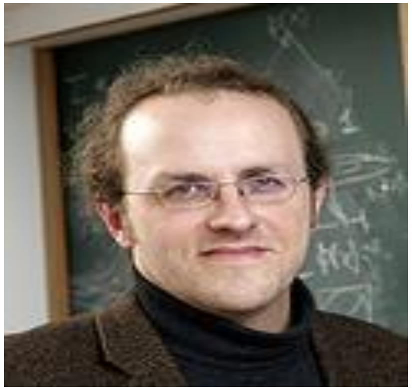
- **General formulation for probabilistic modeling with continuous variables.**

Kenji Fukumizu

Arthur Gretton



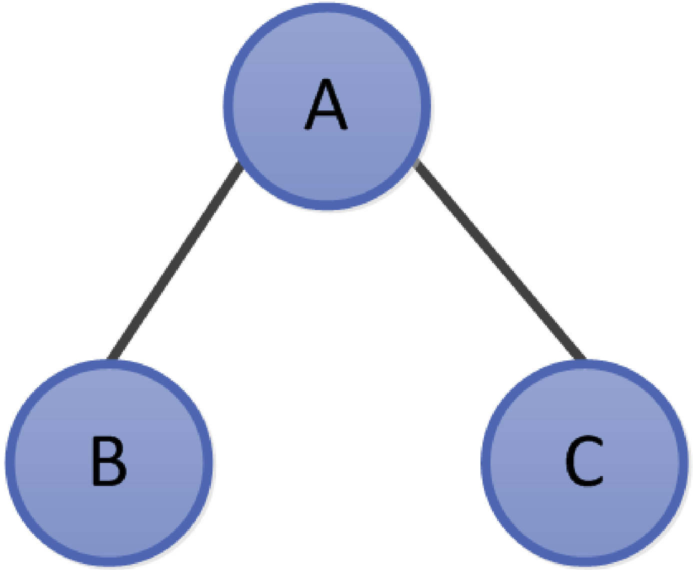
Bernhard Schölkopf



Alex Smola

Le Song

Why do Gaussians Work?



(1) Because we have parameters (sufficient statistics) !!!!

(2) It is easy to marginalize/condition etc.

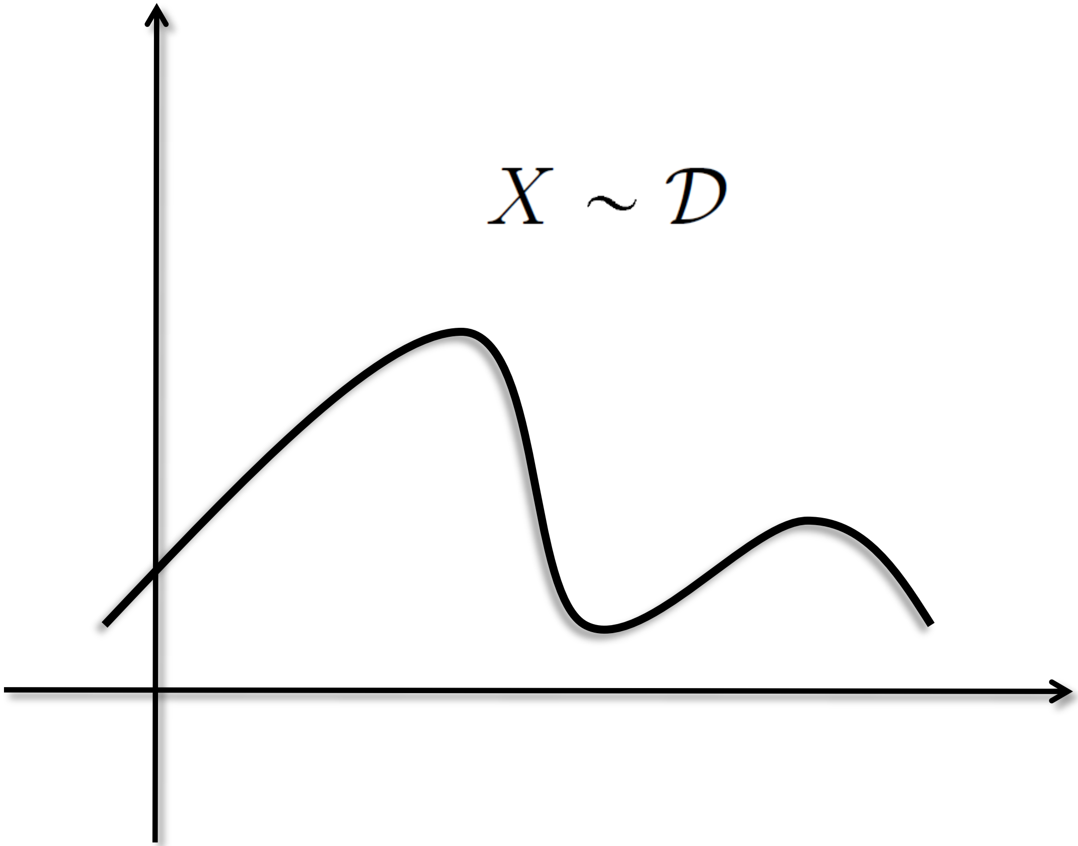
Bijection between (mean,variance) pair and distribution

$$(\mu_1, \sigma_1) \longleftrightarrow N(\mu_1, \sigma_1)$$

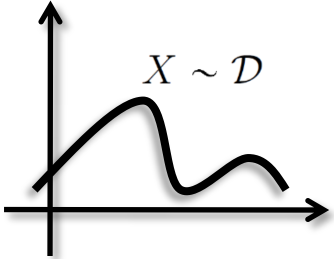
$$(\mu_2, \sigma_2) \longleftrightarrow N(\mu_2, \sigma_2)$$

Key Idea – Create Sufficient Statistic for Arbitrary Distribution

- I want to represent this distribution with a small vector $\boldsymbol{\mu}_{\boldsymbol{X}}$.

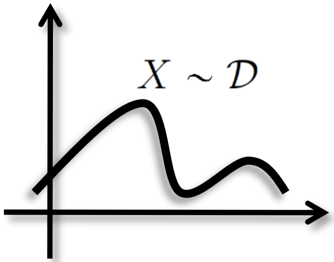


Idea 1: Take some Moments



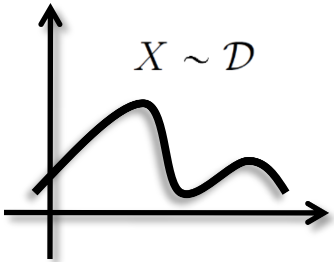
$$\mu_X = \left(\mathbb{E}[X] \right)$$

Problem: Lots of Distributions have the same mean!



$$\mu_X = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \end{pmatrix}$$

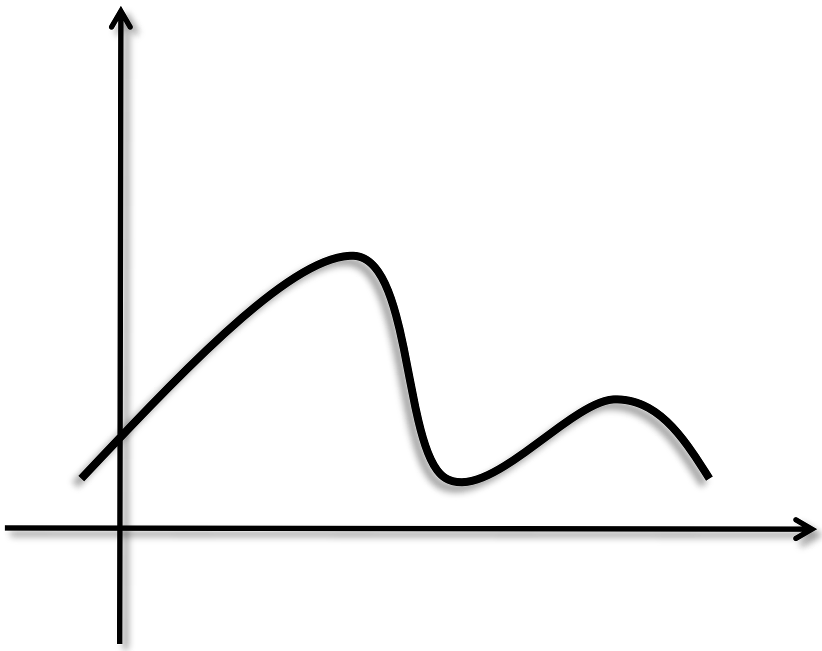
Better, but lots of distributions still have the same mean and variance!



$$\mu_X = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \\ \mathbb{E}[X^3] \end{pmatrix}$$

Even better, but lots of distributions still have the same first three moments!

Better Idea: Create Infinite Dimensional Statistic



$$\boldsymbol{\mu}_X = \begin{pmatrix} \mathbb{E}[X] \\ \mathbb{E}[X^2] \\ \mathbb{E}[X^3] \\ \dots \\ \dots \end{pmatrix}$$

(not exactly, but right idea…)

- But the vector is infinite……..how do we compute things with it?????

Remember the Kernel Trick!!!

Primal Formulation:

$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_j \xi$$

$$(\boldsymbol{w}^\top \boxed{\boldsymbol{\phi}(\boldsymbol{x}_j)} + b)y_j \geqslant 1 - \xi_j \quad \forall j$$

$$\xi_j \geqslant 0 \quad \forall j$$

Infinite, cannot be directly computed

But the dot product is easy to compute ☺

Dual Formulation:

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2}\sum_{i,j} \alpha_i \alpha_j y_i y_j \boxed{\boldsymbol{\phi}(\boldsymbol{x}_i)^\top \boldsymbol{\phi}(\boldsymbol{x}_i)}$$

$$\sum_i \alpha_i y_i = 0$$

$$0 \leqslant \alpha_i \leqslant C \quad \forall i$$

Overview of Hilbert Space Embedding

- Create an infinite dimensional statistic for a distribution.

- Two Requirements:
  - Map from distributions to statistics is one-to-one
  - Although statistic is infinite, it is cleverly constructed such that the kernel trick can be applied.

- Perform Belief Propagation as if these statistics are the conditional probability tables.

- We will now make this construction more formal by introducing the concept of Hilbert Spaces

Vector Space

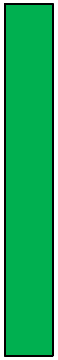
- A set of objects closed under linear combinations:

$$\boldsymbol{v}, \boldsymbol{w} \in \mathcal{V} \implies \alpha \boldsymbol{v} + \beta \boldsymbol{w} \in \mathcal{V}$$

- Normally, you think of these “objects” as finite dimensional vectors. However, in general the objects can be functions.

- **Nonrigorous Intuition**: A function is like an infinite dimensional vector.

$$\boldsymbol{f} =$$



Hilbert Space

- A Hilbert Space is a complete vector space equipped with an inner product.

- The inner product $\langle \boldsymbol{f}, \boldsymbol{g} \rangle$ has the following properties:
  - Symmetry $\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \langle \boldsymbol{g}, \boldsymbol{f} \rangle$
  - Linearity $\langle \alpha \boldsymbol{f}_1 + \beta \boldsymbol{f}_2, \boldsymbol{g} \rangle = \alpha \langle \boldsymbol{f}_1, \boldsymbol{g} \rangle + \beta \langle \boldsymbol{f}_2, \boldsymbol{g} \rangle$
  - Nonnegativity $\langle \boldsymbol{f}, \boldsymbol{f} \rangle \geqslant 0$
  - Zero $\langle \boldsymbol{f}, \boldsymbol{f} \rangle = 0 \implies \boldsymbol{f} = 0$

- Basically a “nice” infinite dimensional vector space, where lots of things behave like the finite case (e.g. using inner product we can define “norm” or “orthogonality”)

Hilbert Space Inner Product

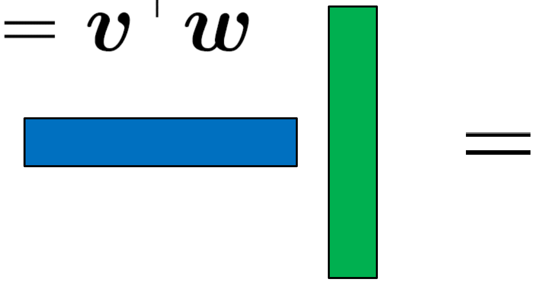
- Example of an inner product (just an example, inner product not required to be an integral)

$$\langle \boldsymbol{f}, \boldsymbol{g} \rangle = \int \boldsymbol{f}(x)\boldsymbol{g}(x)\,dx$$

Inner product of two functions is a number

- **Non-rigorous Intuition**: Like the traditional finite vector space inner product

$$\langle \boldsymbol{v}, \boldsymbol{w} \rangle = \boldsymbol{v}^{\top}\boldsymbol{w}$$



$=$ **scalar**

Linear Operators

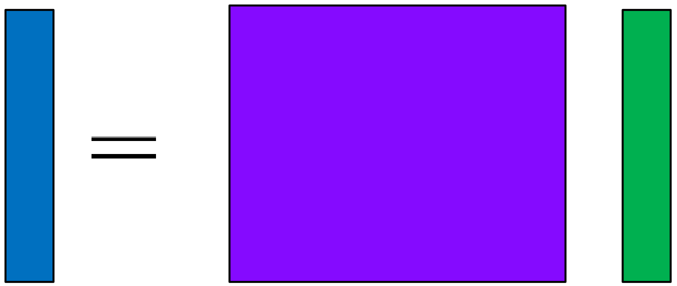
- An operator ***C*** maps a function ***f*** in one Hilbert Space to another function ***g*** in the same or another Hilbert Space.

$$\boldsymbol{g} = \boldsymbol{C}\boldsymbol{f}$$

- Linear Operator:

$$\boldsymbol{C}(\alpha \boldsymbol{f} + \beta \boldsymbol{g}) = \alpha \boldsymbol{C}\boldsymbol{f} + \beta \boldsymbol{C}\boldsymbol{g}$$

- **Non-rigorous Intuition:** Operators are sort of like matrices.



Adjoints (Transposes)

- The adjoint $\mathcal{C}^\top : \mathcal{G} \to \mathcal{F}$ of an operator $\mathcal{C} : \mathcal{F} \to \mathcal{G}$ is defined such that

$$\langle \boldsymbol{g}, \boldsymbol{\mathcal{C}} \boldsymbol{f} \rangle = \langle \boldsymbol{\mathcal{C}}^\top \boldsymbol{g}, \boldsymbol{f} \rangle \qquad \forall \boldsymbol{f} \in \mathcal{F}, \boldsymbol{g} \in \mathcal{G}$$

- Like transpose / conjugate transpose for real / complex matrices:

$$\boldsymbol{w}^\top \boldsymbol{M} \boldsymbol{v} = (\boldsymbol{M}^\top \boldsymbol{w})^\top \boldsymbol{v}$$

Hilbert Space Outer Product

$\boldsymbol{f} \otimes \boldsymbol{g}$ is implicitly defined such that

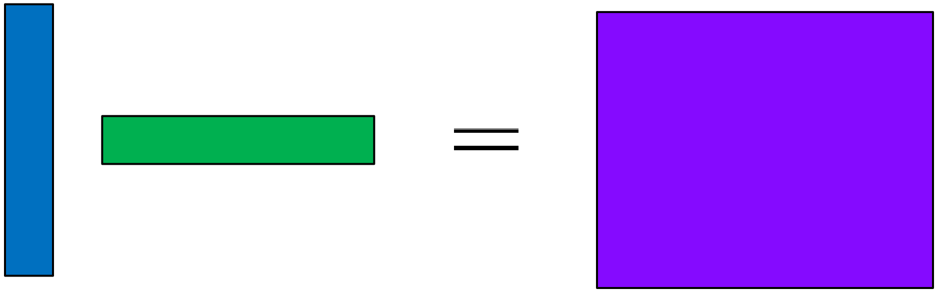
$$\boldsymbol{f} \otimes \boldsymbol{g}(\boldsymbol{h}) = \langle \boldsymbol{g}, \boldsymbol{h} \rangle \boldsymbol{f}$$

Outer Product of two functions is an operator

- **Non-rigorous Intuition:** Like Vector Space Outer Product

$$\boldsymbol{v} \otimes \boldsymbol{w} = \boldsymbol{v}\boldsymbol{w}^\top$$

$$\boldsymbol{v}\boldsymbol{w}^\top(\boldsymbol{z}) = \langle \boldsymbol{w}, \boldsymbol{z} \rangle \boldsymbol{v}$$



Reproducing Kernel Hilbert Space

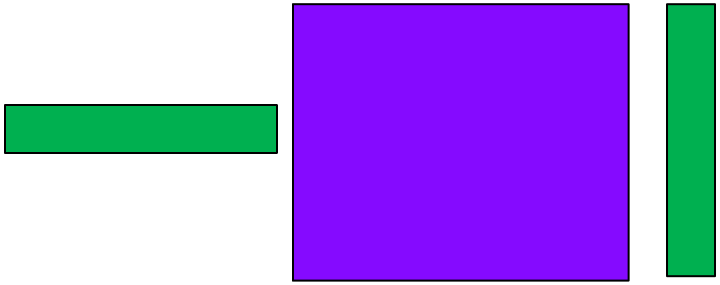
- Basically, a "really nice" infinite dimensional vector space where even more things behave like the finite case

- We are going to "construct" our Reproducing Kernel Hilbert Space with a **Mercer Kernel**. A Mercer Kernel $\boldsymbol{K}(\boldsymbol{x},\boldsymbol{y})$ is a function of two variables, such that:

$$\int\int \boldsymbol{K}(x,y)\boldsymbol{f}(x)\boldsymbol{f}(y)\,dx\,dy > 0 \quad \forall \boldsymbol{f}$$

- The is a generalization of a positive definite matrix:

$$\boldsymbol{x}^\top \boldsymbol{A}\boldsymbol{x} > 0 \quad \forall \boldsymbol{x}$$



$$> 0$$

Gaussian Kernel

- The most common kernel that we will use is the Gaussian RBF Kernel:

$$\boldsymbol{K}(x, y) = \exp\left(\frac{\|x - y\|_2^2}{\sigma^2}\right)$$

The Feature Function

- Consider holding one element of the kernel fixed. We get a function of one variable which we call the feature function. The collection of feature functions is called the **feature map**.

$$\phi_x := \boldsymbol{K}(x, \cdot)$$

- For a Gaussian Kernel the feature functions are unnormalized Gaussians:

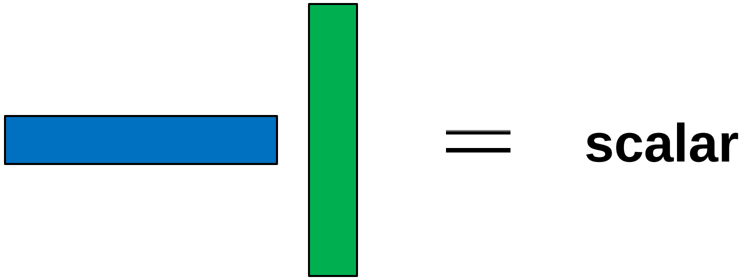
$$\phi_1(y) = \exp\left(\frac{\|1 - y\|_2^2}{\sigma^2}\right)$$

$$\phi_{1.5}(y) = \exp\left(\frac{\|1.5 - y\|_2^2}{\sigma^2}\right)$$

Defining the Inner Product

- Define the Inner Product as:

$$\langle \phi_x, \phi_y \rangle = \langle \boldsymbol{K}(x,\cdot), \boldsymbol{K}(y,\cdot) \rangle := \boldsymbol{K}(x,y)$$



= **scalar**

- Note that:

$$\phi_x(y) = \phi_y(x) = \boldsymbol{K}(x,y)$$

Reproducing Kernel Hilbert Space

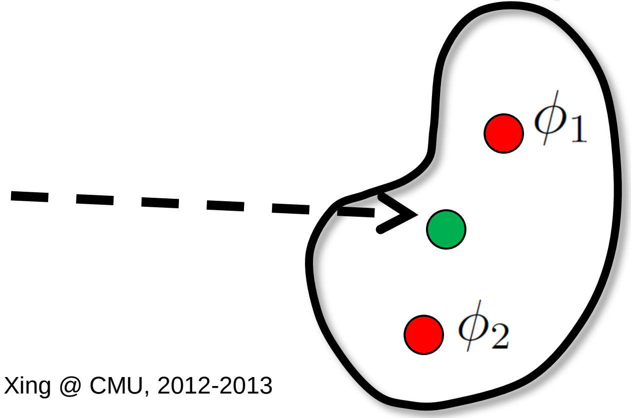
- Consider the set of functions that can be formed with linear combinations of these feature functions:

$$\mathcal{F}_0 = \left\{ f(z) : \sum_{j=1}^{k} \alpha_j \phi_{x_j}(z), \forall k \in \mathbb{N}_+ \text{ and } x_j \in \mathcal{X} \right\}$$

- We define the Reproducing Kernel Hilbert Space $\mathcal{F}$ to the completion of $\mathcal{F}_0$ (like $\mathcal{F}_0$ with the "holes" filled in)

- Intuitively, the feature functions are like an over-complete basis for the RKHS

$$\boldsymbol{f}(z) = \alpha_1 \phi_1(z) + \alpha_2 \phi_2(z)$$



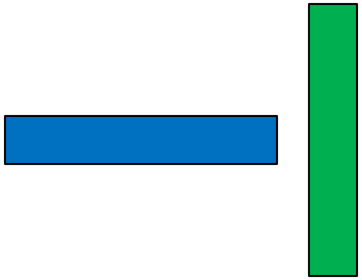
Reproducing Property

- It can now be derived that the inner product of a function **f** with $\boldsymbol{\phi}_X$, evaluates a function at point ***x***:

$$\begin{aligned}\langle \boldsymbol{f}, \boldsymbol{\phi}_x \rangle &= \left\langle \sum_j \alpha_j \boldsymbol{\phi}_{x_j}, \boldsymbol{\phi}_x \right\rangle \\ &= \sum_j \alpha_j \langle \boldsymbol{\phi}_{x_j}, \boldsymbol{\phi}_x \rangle && \textbf{Linearity of inner product} \\ &= \sum_j \alpha_j \boldsymbol{K}(x_j, x) && \textbf{Definition of kernel} \\ &= \boldsymbol{f}(x)\end{aligned}$$

Remember that

$$\boldsymbol{K}(x_j, x) := \boldsymbol{\phi}_{x_j}(x)$$



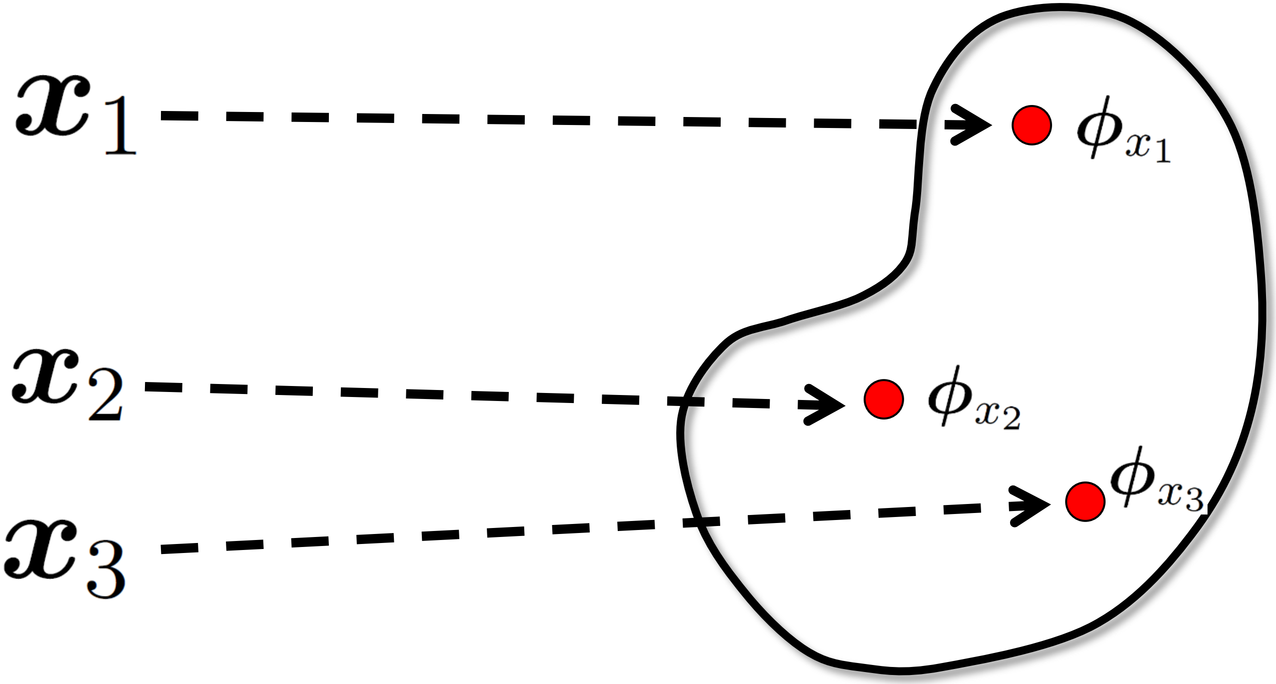
$=$ **scalar**

SVM Kernel Intuition

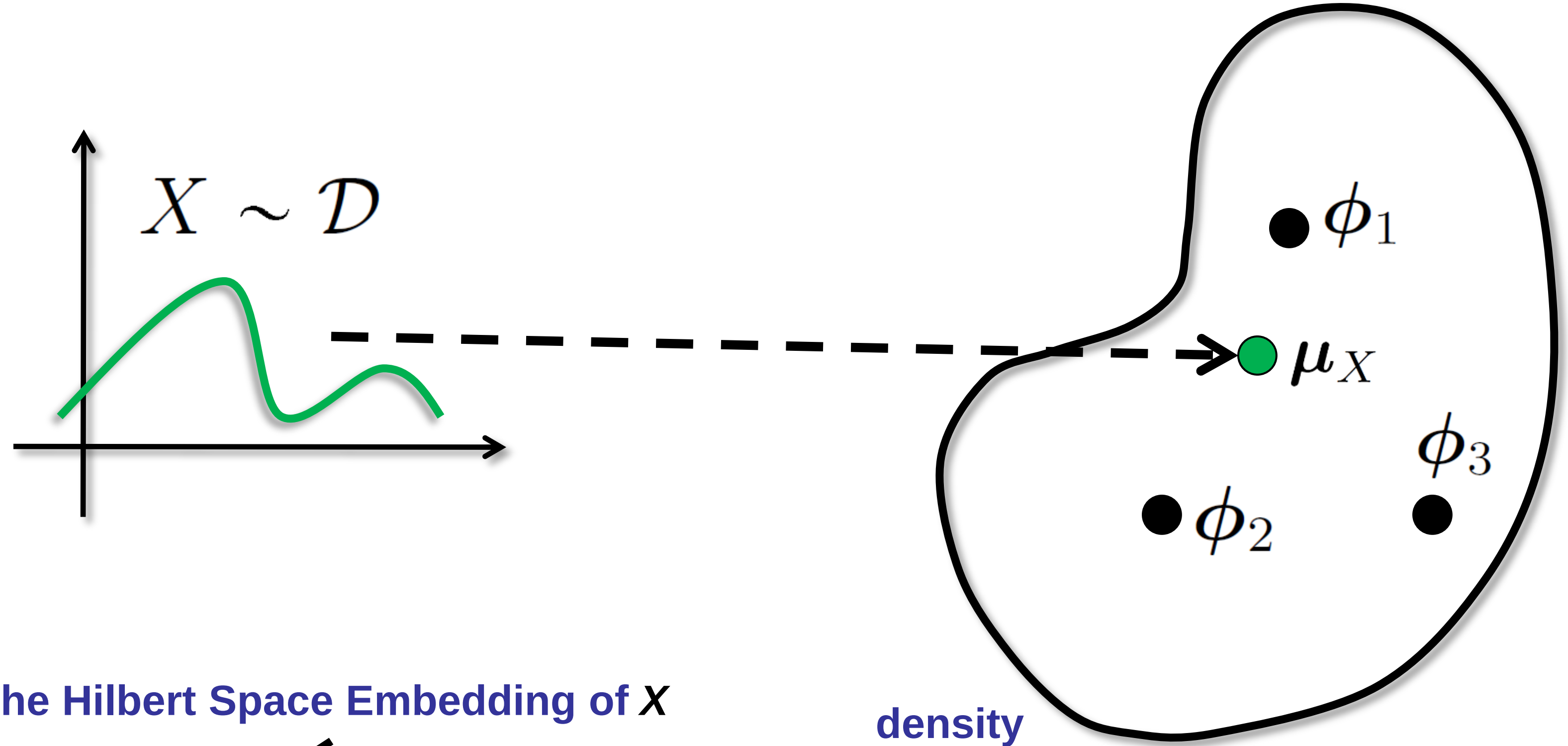
$$\min_{\boldsymbol{w},b} \frac{1}{2}\boldsymbol{w}^\top \boldsymbol{w} + C\sum_j \xi$$

$$(\boldsymbol{w}^\top \phi(\boldsymbol{x}_j) + b)y_j \geqslant 1 - \xi_j \quad \forall j \qquad \xi_j \geqslant 0 \quad \forall j$$

Maps features to RKHS Feature Functions!



How To Embed Distributions (Mean Map) [Smola et al. 2007]



The Hilbert Space Embedding of *X*

density

$$\boldsymbol{\mu}_X(\cdot) = \mathbb{E}_{X\sim\mathcal{D}}[\phi_X] = \int \boldsymbol{p}_{\mathcal{D}}(X)\phi_X(\cdot)\,dX$$

Mean Map cont.

- Mean Map

$$\boldsymbol{\mu}_X = \mathbb{E}_X[\phi_X]$$

- If the kernel is universal, then the map from distributions to embeddings is one-to-one. Examples of universal kernels:
  - Gaussian RBF Kernel.
  - Laplace Kernel

- “Empirical Estimate” (not actually computable from data if feature map is infinite….but we will solve this problem in the next lecture)

$$\hat{\boldsymbol{\mu}}_X = \frac{1}{N}\sum_{n=1}^{N}\phi_{x_n}$$

Data point

Example (Discrete)

- Consider a random variable ***X*** that takes the values $\mathbf{1, 2, 3, 4}$.

  We want to embed it into an ***RKHS***. Which ***RKHS***?

- The RKHS of 4 dimensional vectors in $\boldsymbol{R}^{\mathbf{4}}$. The feature functions in this RKHS are:

$$\phi_1 = \begin{pmatrix} 1 \\ 0 \\ 0 \\ 0 \end{pmatrix} \phi_2 = \begin{pmatrix} 0 \\ 1 \\ 0 \\ 0 \end{pmatrix} \phi_3 = \begin{pmatrix} 0 \\ 0 \\ 1 \\ 0 \end{pmatrix} \phi_4 = \begin{pmatrix} 0 \\ 0 \\ 0 \\ 1 \end{pmatrix}$$

$$\mu_X = \mathbb{E}_X[\phi_X] = \mathbb{P}[X = 1]\phi_1 + \mathbb{P}[X = 2]\phi_2 + \mathbb{P}[X = 3]\phi_3 + \mathbb{P}[X = 4]\phi_4$$

$$\mu_X = \begin{pmatrix} \mathbb{P}[X = 1] \\ \mathbb{P}[X = 2] \\ \mathbb{P}[X = 3] \\ \mathbb{P}[X = 4] \end{pmatrix}$$

Embedding equal to marginal probability vector in the discrete case

Mean Map cont.

$$\mathbb{E}_{X\sim\mathcal{D}}[\boldsymbol{f}(X)] = \langle \boldsymbol{f}, \boldsymbol{\mu}_X \rangle$$ If ***f*** is in the RKHS

- Why?

$$\begin{aligned}\langle \boldsymbol{f}, \boldsymbol{\mu}_X \rangle &= \langle \boldsymbol{f}, \mathbb{E}_{X\sim\mathcal{D}}[\phi_X] \rangle \\ &= \mathbb{E}_{X\sim\mathcal{D}}[\boldsymbol{f}(X)]\end{aligned}$$

Embedding Joint Distribution of 2 Variables [Smola et al. 2007]

- Define the uncentered cross-covariance operator $\boldsymbol{C_{YX}}$ implicitly such that

$$\langle \boldsymbol{g}, \boldsymbol{\mathcal{C}}_{YX} \boldsymbol{f} \rangle = \mathbb{E}_{YX}[\boldsymbol{f}(X)\boldsymbol{g}(Y)] \quad \forall \boldsymbol{f} \in \mathcal{F}, \forall \boldsymbol{g} \in \mathcal{G}$$

- Note how $\boldsymbol{f}$ is in one Hilbert Space, while $\boldsymbol{g}$ is in another.

- $\boldsymbol{C_{YX}}$ will be our embedding of the joint distribution of $\boldsymbol{X}$ and $\boldsymbol{Y}$.

- Note how $\boldsymbol{C_{YX}}$ is an operator, just like $\boldsymbol{P}[\boldsymbol{X}, \boldsymbol{Y}]$ is a matrix.

Cross Covariance Operator cont.

- Let $\phi_X \in \mathcal{F}$ and $\psi_Y \in \mathcal{G}$ (the feature functions of these two RKHSs)

- Then explicit form of cross-covariance operator is:

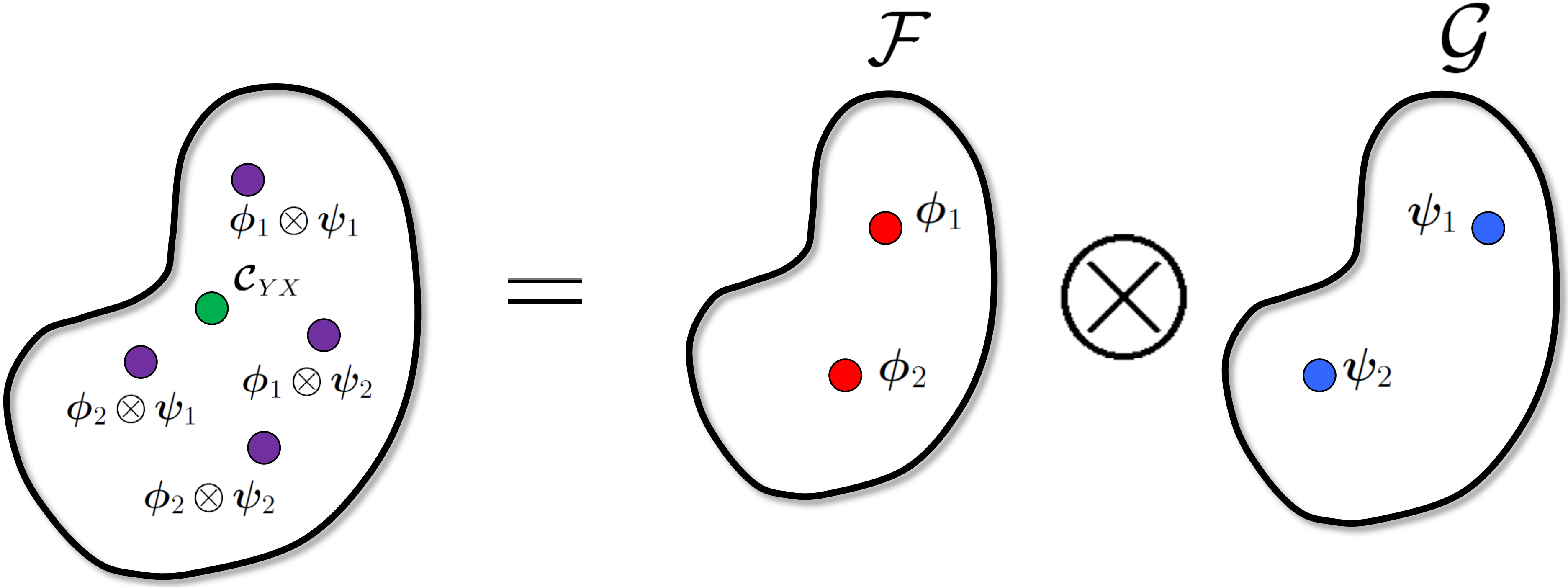
$$\mathcal{C}_{YX} = \mathbb{E}_{YX}[\psi_Y \otimes \phi_X]$$

- Looks like the Uncentered Covariance of two variables ***X*** and ***Y***:

$$\mathrm{Cov}(X, Y) = \mathbb{E}_{YX}[YX]$$

Embedding Joint Distribution of Two Variables [Smola et al. 2007]

Embed in the Tensor Product of two RKHS's



“Tensor Product” Intuition

- Consider two finite sets:

$$\boldsymbol{S} = \{1, 3, 4\} \qquad\qquad \boldsymbol{T} = \{2, 6\}$$

- If “outer product” is defined as:

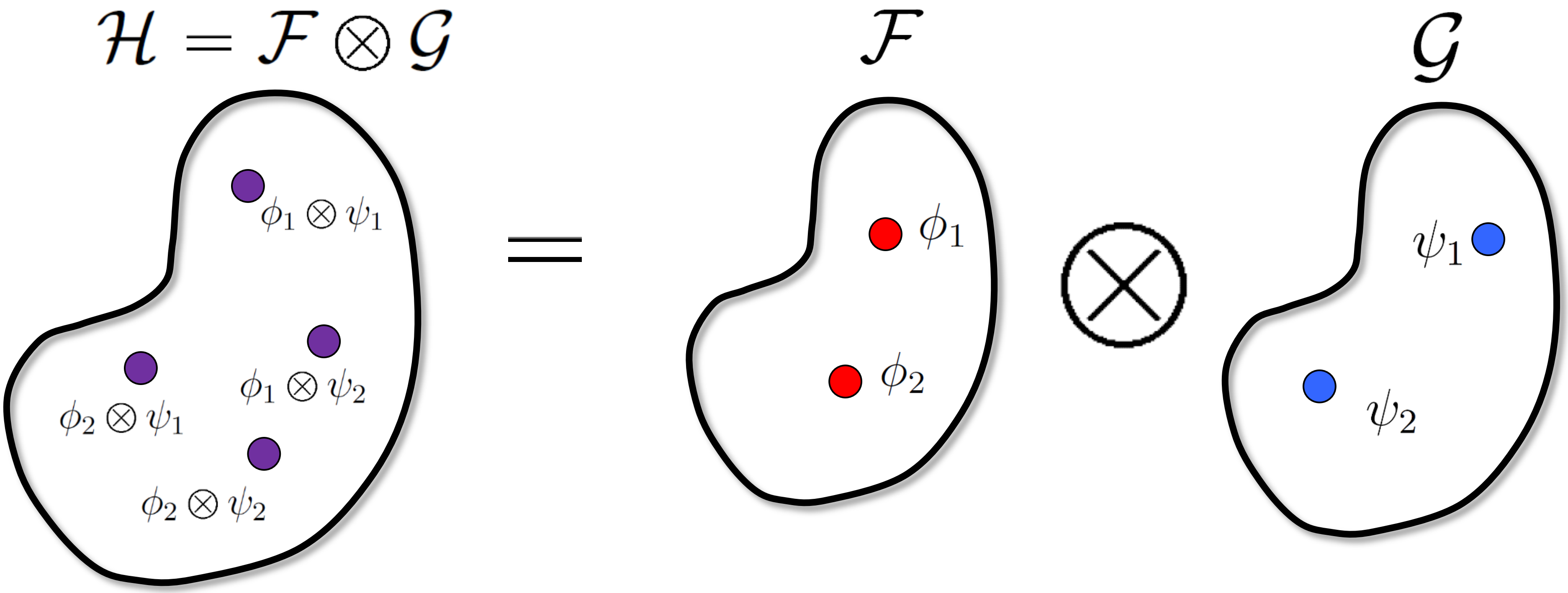
$$a \otimes b = (a, b)$$

- Then tensor product is:

$$\boldsymbol{S} \otimes \boldsymbol{T} = \{(1,2),(1,6),(3,2),(3,6),(4,3),(4,6)\}$$

- (Don’t take the example too literally since this is not a vector space)

Tensor Product of Two Vector Spaces



$$\mathcal{H} = \{\boldsymbol{h} : \exists \boldsymbol{f} \in \mathcal{F}, \boldsymbol{g} \in \mathcal{G} \text{ s.t. } \boldsymbol{h} = \boldsymbol{f} \otimes \boldsymbol{g}\}$$

Cross Covariance Operator cont.

- **Proof:**

$$
\begin{aligned}
\langle \boldsymbol{g}, \boldsymbol{\mathcal{C}}_{YX}\boldsymbol{f} \rangle &= \langle \boldsymbol{g}, \mathbb{E}_{YX}[\psi_Y \otimes \phi_X]\boldsymbol{f} \rangle \\
&= \mathbb{E}_{YX}\left[\langle \boldsymbol{g}, [\psi_Y \otimes \phi_X]\boldsymbol{f} \rangle\right] && \text{Move expectation outside} \\
&= \mathbb{E}_{YX}\left[\langle \boldsymbol{g}, \langle \phi_X, \boldsymbol{f} \rangle \psi_Y \rangle\right] && \text{Definition of outer product} \\
&= \mathbb{E}_{YX}\left[\langle \boldsymbol{g}, \psi_Y \rangle \langle \boldsymbol{f}, \phi_X \rangle\right] && \text{Rearrange} \\
&= \mathbb{E}_{YX}\left[\boldsymbol{g}(Y)\boldsymbol{f}(X)\right] && \text{Reproducing Property}
\end{aligned}
$$

Auto Covariance Operator

- The uncentered auto-covariance operator is:

$$\mathcal{C}_{XX} = \mathbb{E}_X[\phi_X \otimes \phi_X]$$

- Looks like the uncentered variance of X

$$\text{Uncentered-Var}(X) = \mathbb{E}[X^2]$$

- **Intuition:** Analogous to

$$\text{Diag}(\mathbb{P}[X])$$

Conditional Embedding Operator

- Conditional Embedding Operator:

$$\mathcal{C}_{\mathbf{Y}|\mathbf{X}} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

- Intuition:

$$\mathbb{P}[Y|X] = \mathbb{P}[Y, X] \times \mathrm{Diag}(\mathbb{P}[X])^{-1}$$

Conditional Embedding Cont.

- Conditional Embedding Operator:

$$\mathcal{C}_{\boldsymbol{Y}|\boldsymbol{X}} = \mathcal{C}_{YX}\mathcal{C}_{XX}^{-1}$$

- Has Following Property:

$$\mathbb{E}_{Y|x}[\phi_Y|x] = \mathcal{C}_{Y|X}\phi_x$$

- Analogous to “Slicing” a Conditional Probability Table in the Discrete Case:

$$\mathbb{P}[Y|X=1] = \mathbb{P}[Y|X]\boldsymbol{\delta}_1$$

Why We Care

- So we have some statistics for marginal, joint, and conditional distributions….
- How does this help us define Belief Propagation?
- There are many parametric distributions where it is hard to define message passing

- Think Back: What makes Gaussians different?
  - Easy to marginalize, perform Chain Rule with Gaussians!

Why we Like Hilbert Space Embeddings

We can marginalize and use chain rule in Hilbert Space too!!!

Sum Rule:

$$\mathbb{P}[X] = \int_Y \mathbb{P}[X,Y] = \int_Y \mathbb{P}[X|Y]\mathbb{P}[Y]$$

Sum Rule in RKHS:

$$\boldsymbol{\mu}_X = \mathcal{C}_{X|Y}\boldsymbol{\mu}_Y$$

Chain Rule:

$$\mathbb{P}[X,Y] = \mathbb{P}[X|Y]\mathbb{P}[Y] = \mathbb{P}[Y|X]\mathbb{P}[Y]$$

Chain Rule in RKHS:

$$\mathcal{C}_{YX} = \mathcal{C}_{Y|X}\mathcal{C}_{XX} = \mathcal{C}_{X|Y}\mathcal{C}_{YY}$$

We will prove these in the next lecture

Summary

- Hilbert Space Embedding provides a way to create a “sufficient statistic” for an arbitrary distribution.

- Can embed marginal, joint, and conditional distributions into the RKHS

- **Next time:**
  - Prove sum rule and chain rule for RKHS embedding
  - Performing Belief Propagation with the Embedding Operators
  - Why the messages are easily computed from data (and not infinite)

References

- Smola, A. J., Gretton, A., Song, L., and Schölkopf, B., A Hilbert Space Embedding for Distributions, Algorithmic Learning Theory, E. Takimoto (Eds.), Lecture Notes on Computer Science, Springer, 2007.

- L. Song. Learning via Hilbert space embedding of distributions. PhD Thesis 2008.

- Song, L., Huang, J., Smola, A., and Fukumizu, K., Hilbert space embeddings of conditional distributions, International Conference on Machine Learning, 2009.