

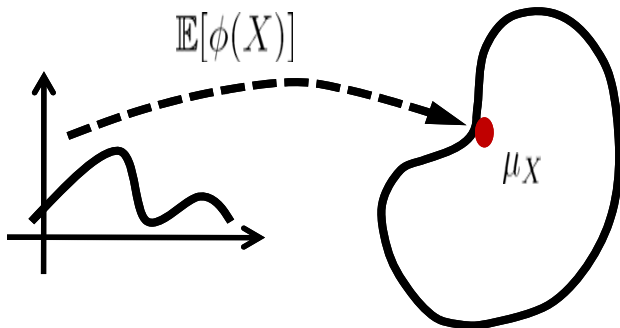
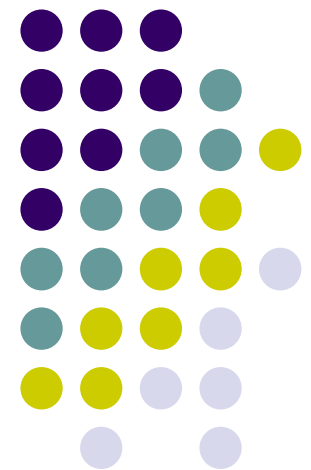


Probabilistic Graphical Models

Introduction to Hilbert Space Embeddings and Kernel GM

Eric Xing

Lecture 22, April 4, 2016



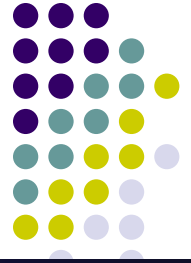
Acknowledgement: slides first drafted by Ankur Parikh

The Optimization View of Graphical Models



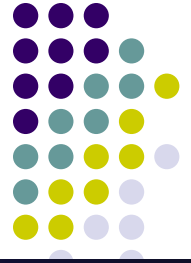
- The connection between optimization and graphical models has led to many amazing discoveries
 - EM
 - Variational Inference
 - Max Margin/Max Entropy Learning
 - Bridge to Statistical Physics, Numerical Methods Communities
- Optimization has many advantages:
 - It is easy to formulate
 - Can derive 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

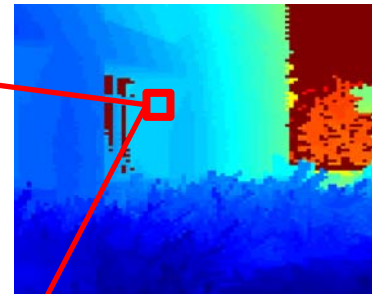
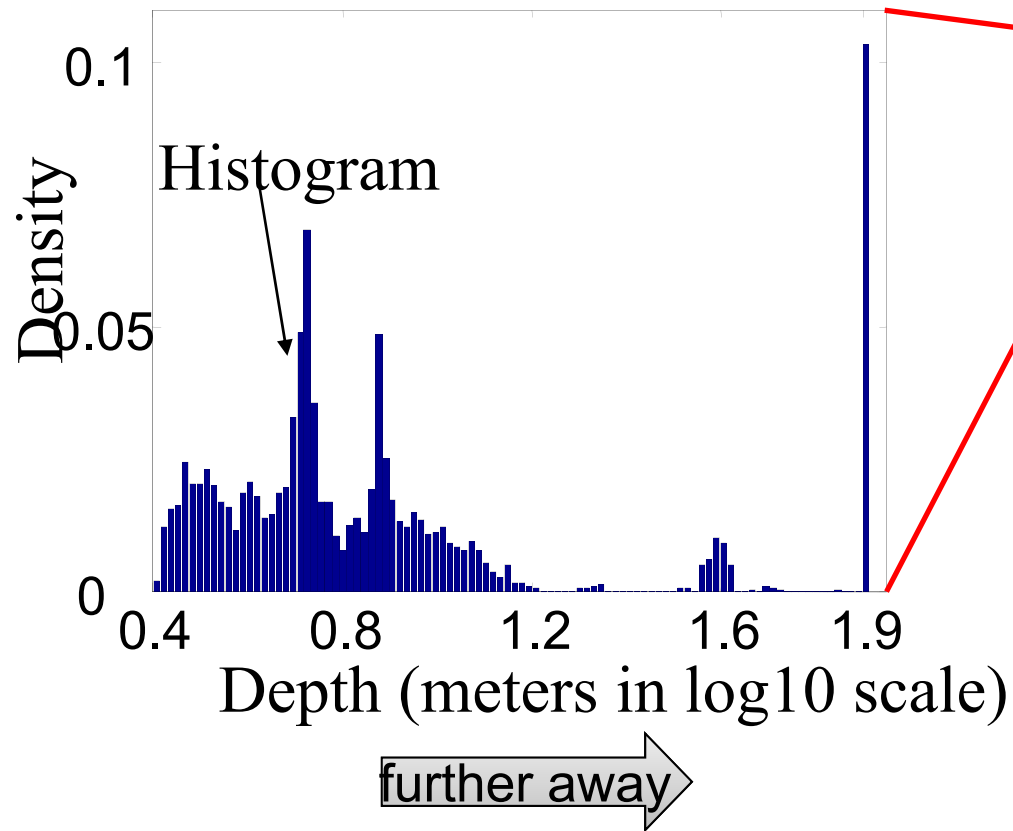


- We are going to discuss a different (still not fully understood) point of view of graphical models that revolves around linear algebra.
- Compared to the optimization perspective, the linear algebra view often less intuitive to formulate.
- However, it lets us solve problems that are intractable from 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

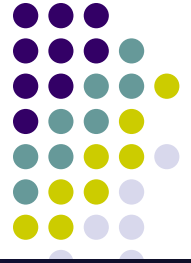


Predict

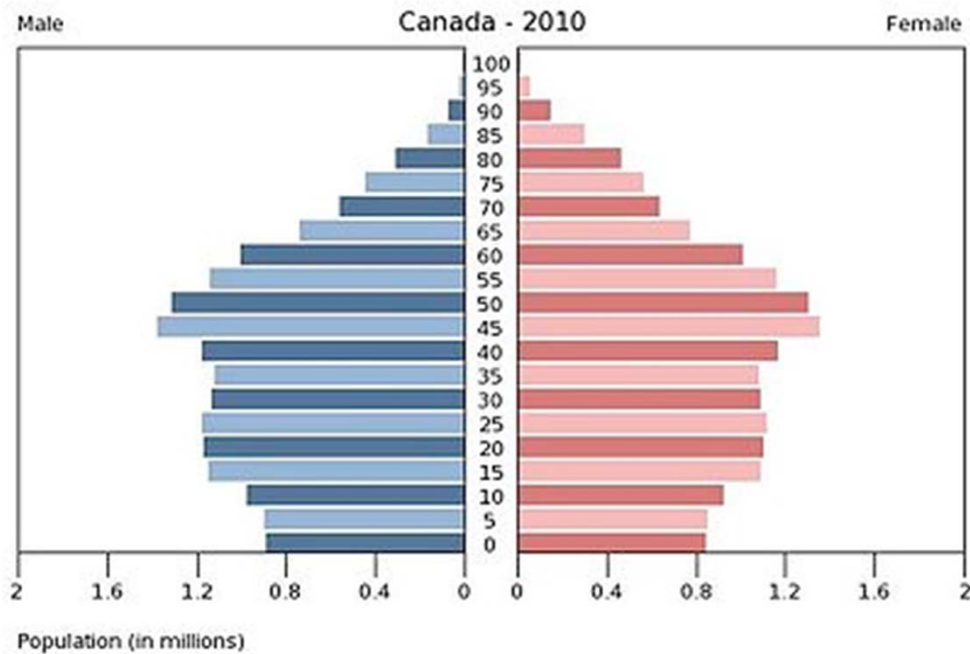


[Saxena, Chung and Ng 2005]

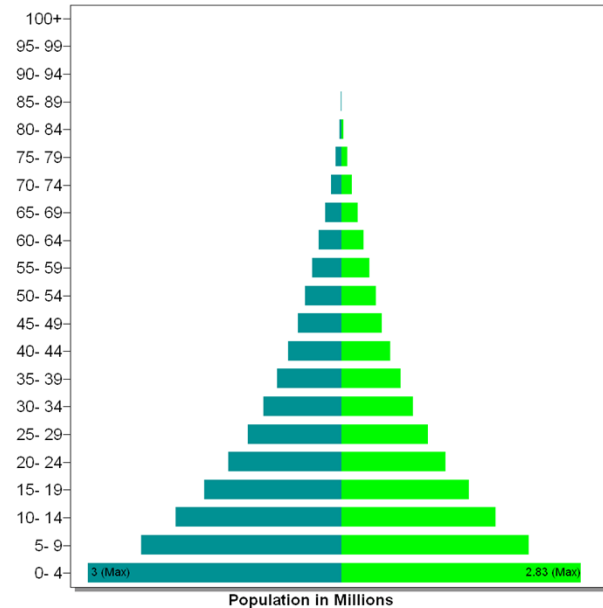
Non-Gaussian Continuous Variables



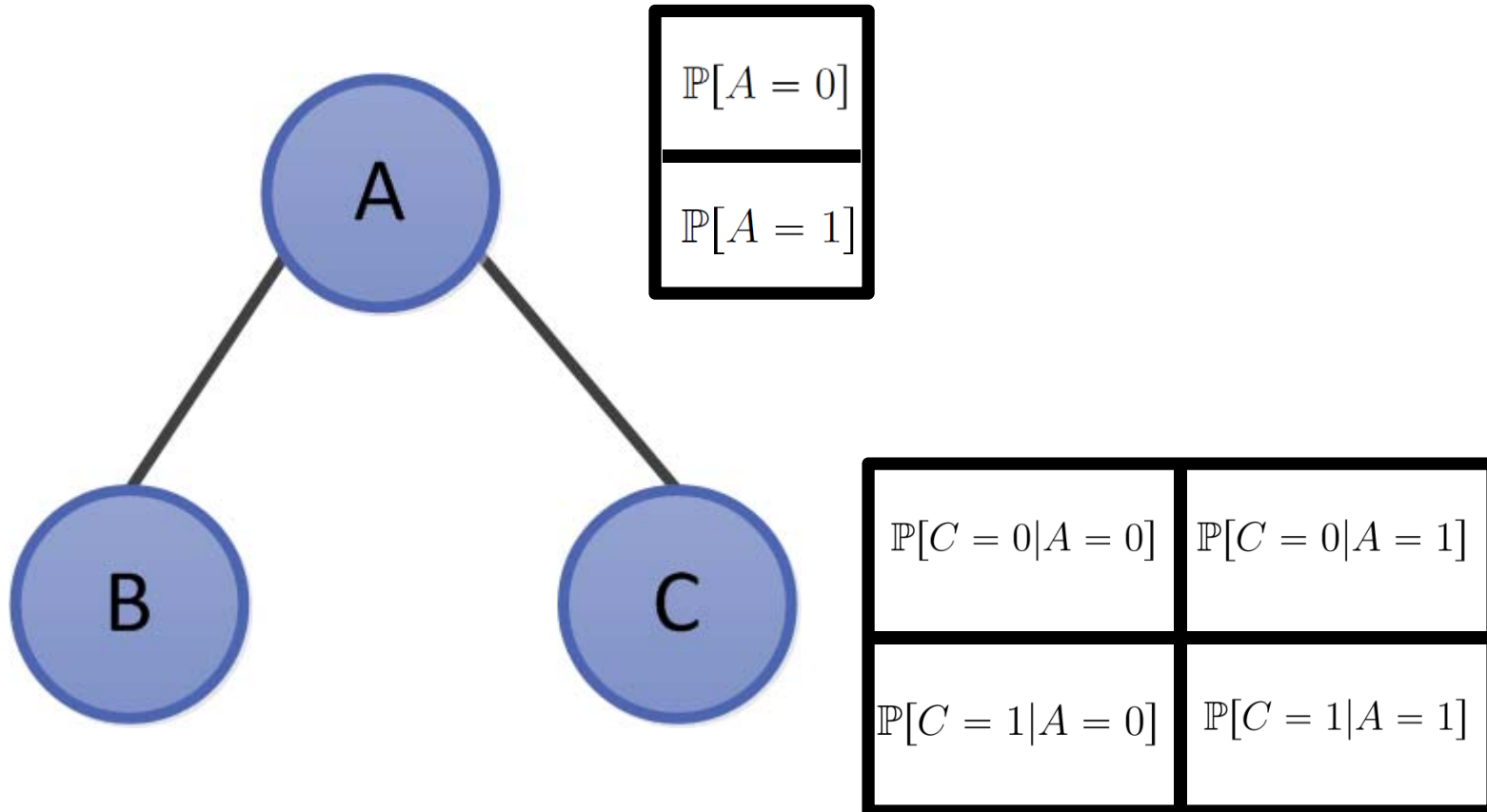
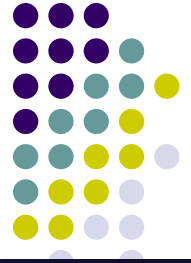
Demographics: Model relationships among different demographic variables



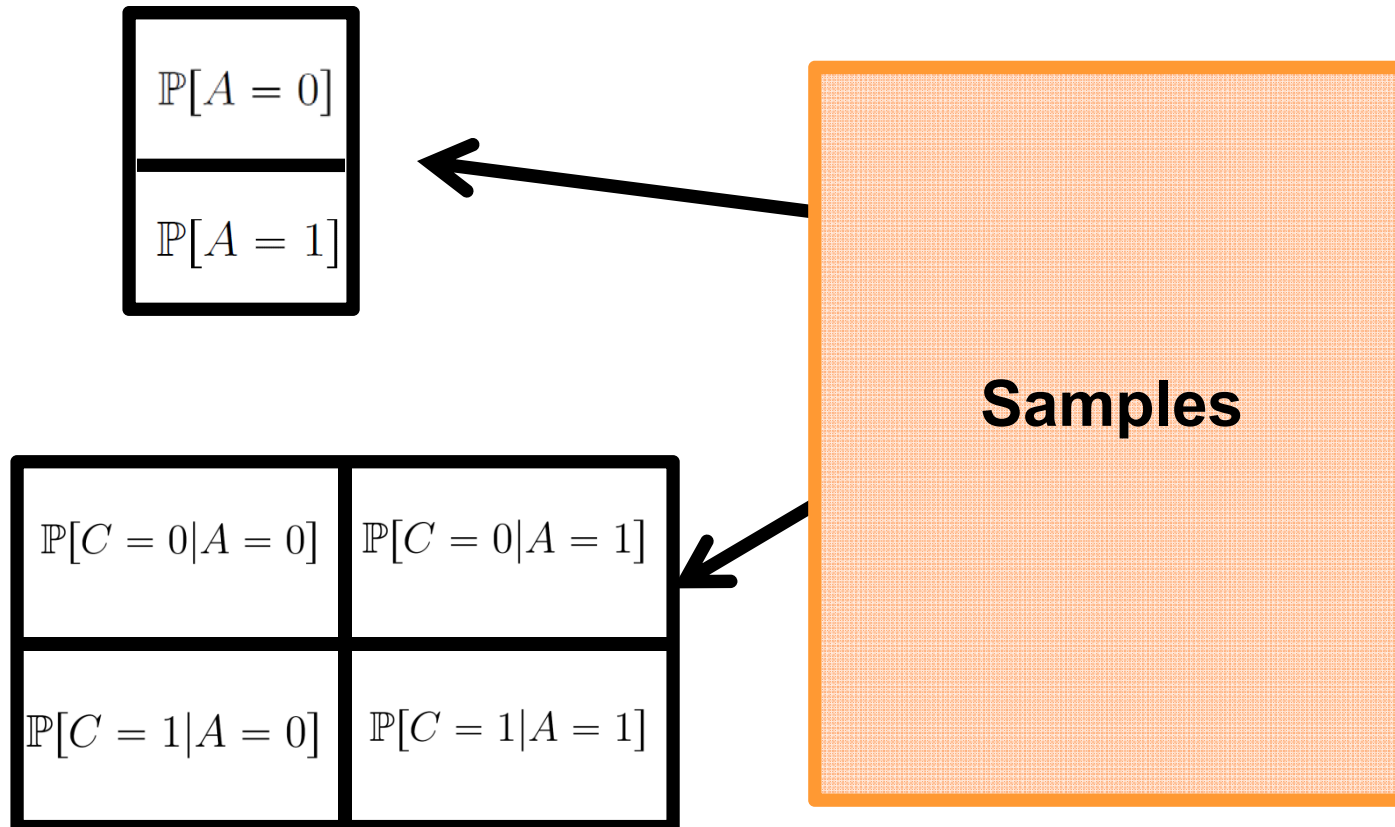
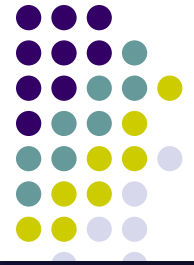
Population Distribution for Afghanistan in Year 2005 [Base Case]



Graphical Models - What we have learned so far...

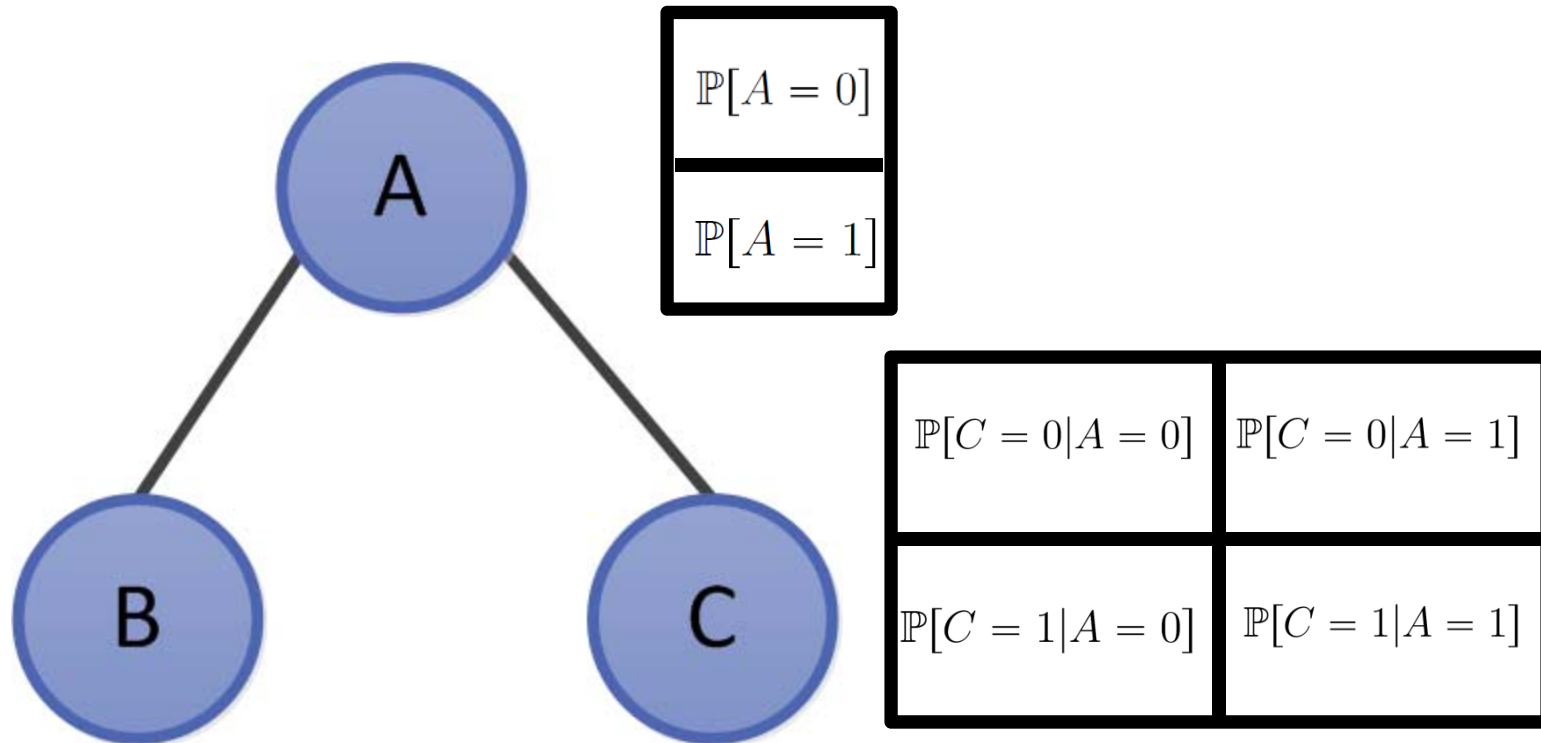
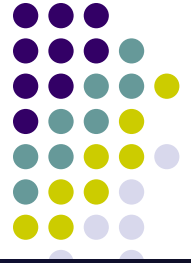


Parameter Learning - What we have learned so far...



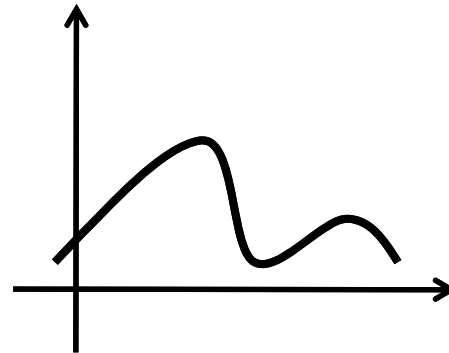
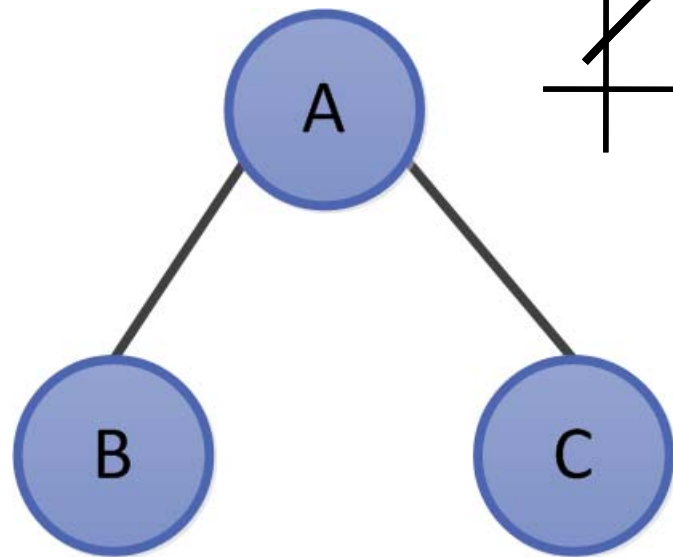
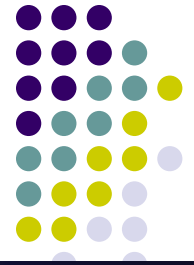
- If variables are observed, just count from dataset
- In case of hidden variables, can use Expectation Maximization.....

Inference - What we have learned so far...

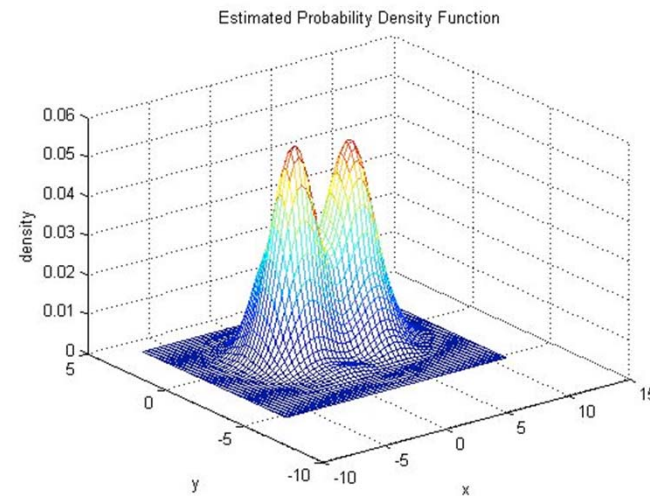


- Can do exact inference with Variable Elimination, Belief Propagation.
- Can do approximate inference with Loopy BP, Mean Field, MCMC

Non-Parametric Continuous Case is Much Harder...

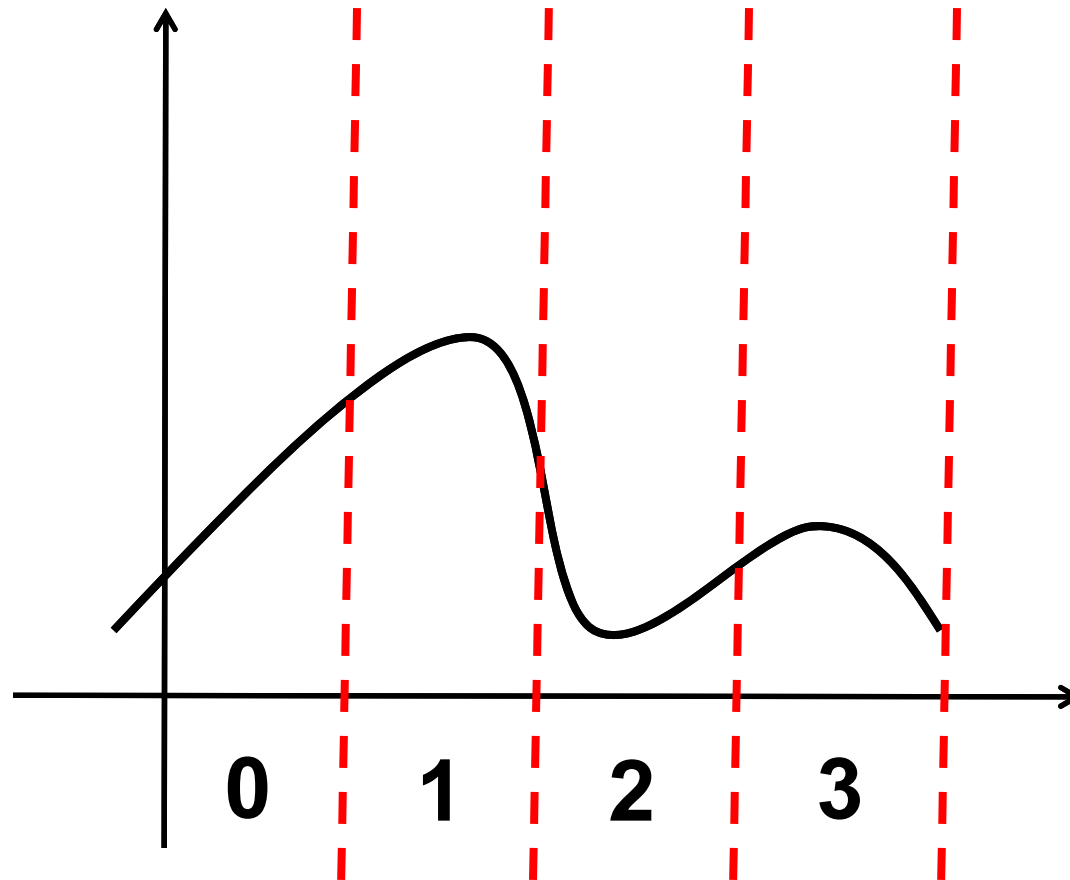


How do we make a
conditional probability
table out of this?



- **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

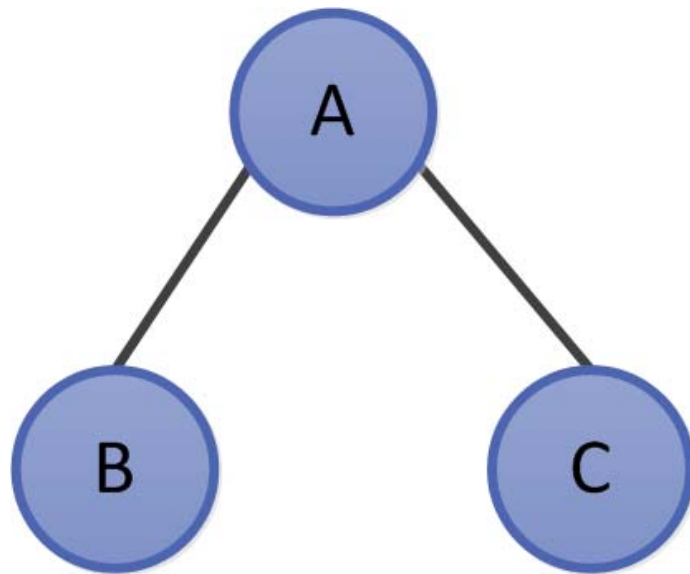


Le Song





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

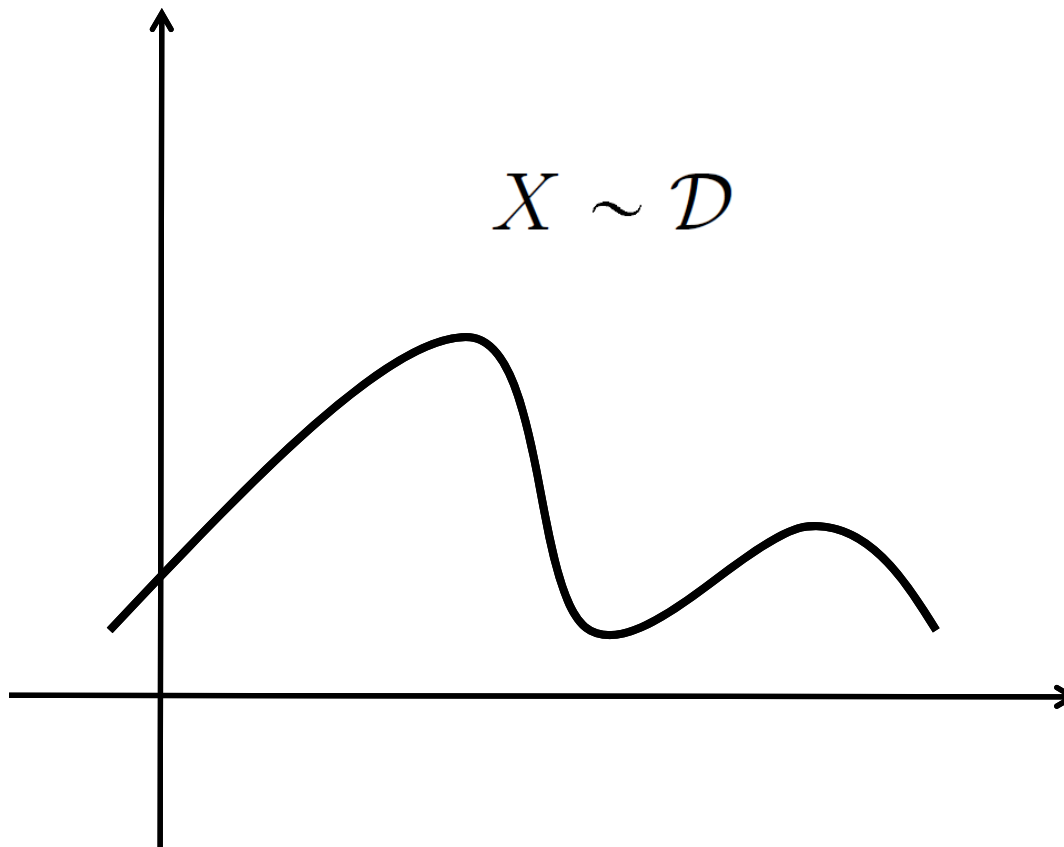
$$(\mu_1, \sigma_1) \longleftrightarrow N(\mu_1, \sigma_1)$$

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

Key Idea – Create Sufficient Statistic for Arbitrary Distribution

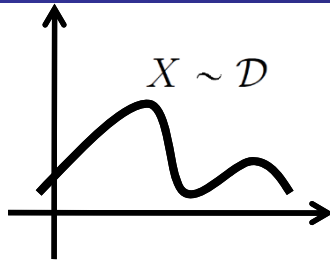


- I want to represent this distribution with a small vector μ_X .



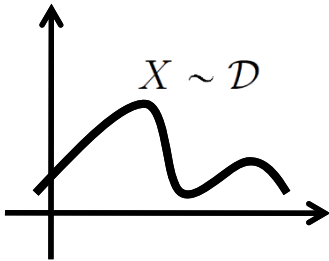


Idea 1: Take some Moments



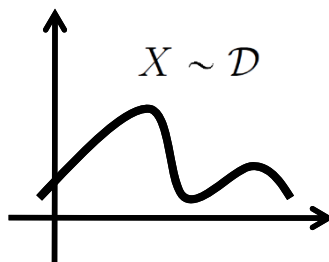
$$\mu_X = \left(\mathbb{E}[X] \right)$$

Problem: Lots of Distributions have the same mean!



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

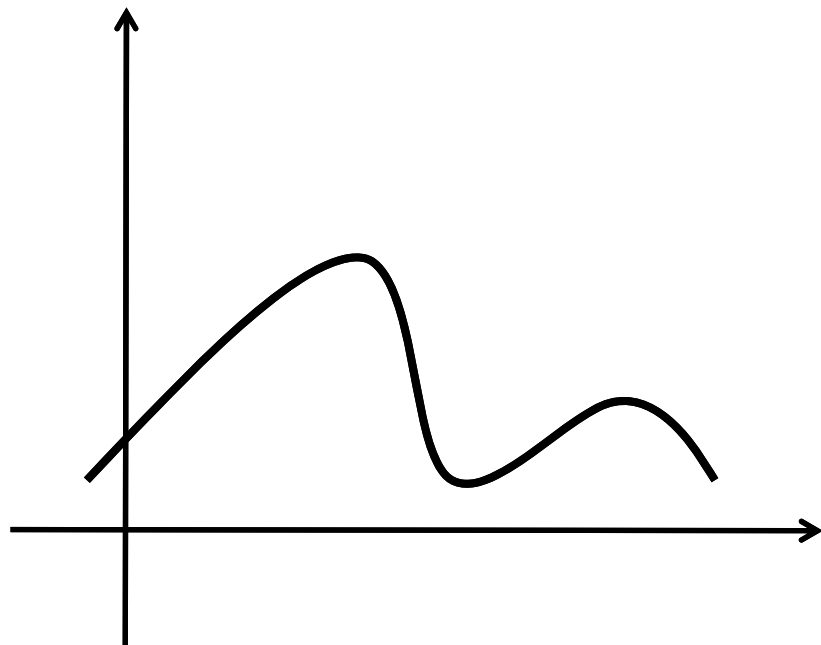
Better, but lots of distributions still have the same mean and variance!



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

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

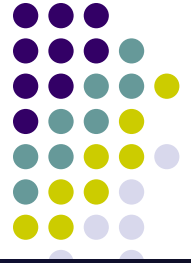
Better Idea: Create Infinite Dimensional Statistic



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

(not exactly, but right idea...)

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



Remember the Kernel Trick!!!

Primal Formulation:

$$\min_{w,b} \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_j \xi_j$$
$$(\mathbf{w}^\top \phi(\mathbf{x}_j) + b)y_j \geq 1 - \xi_j \quad \forall j$$
$$\xi_j \geq 0 \quad \forall j$$

Infinite, cannot be directly computed

But the dot product is easy to compute ☺

Dual Formulation:

$$\max_{\alpha} \sum_i \alpha_i - \frac{1}{2} \sum_{i,j} \alpha_i \alpha_j y_i y_j \phi(\mathbf{x}_i)^\top \phi(\mathbf{x}_j)$$
$$\sum_i \alpha_i y_i = 0$$
$$0 \leq \alpha_i \leq C \quad \forall i$$

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:

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

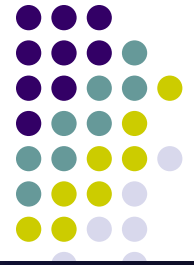
- Normally, you think of these “objects” as finite dimensional vectors. However, in general the objects can be functions.
- **Nonrigorous Intuition:** A function is like an infinite dimensional vector.

$$f = \begin{array}{|c} \hline \text{ } \\ \hline \end{array}$$



Hilbert Space

- A Hilbert Space is a complete vector space equipped with an inner product.
- The inner product $\langle \mathbf{f}, \mathbf{g} \rangle$ has the following properties:
 - Symmetry $\langle \mathbf{f}, \mathbf{g} \rangle = \langle \mathbf{g}, \mathbf{f} \rangle$
 - Linearity $\langle \alpha \mathbf{f}_1 + \beta \mathbf{f}_2, \mathbf{g} \rangle = \alpha \langle \mathbf{f}_1, \mathbf{g} \rangle + \beta \langle \mathbf{f}_2, \mathbf{g} \rangle$
 - Nonnegativity $\langle \mathbf{f}, \mathbf{f} \rangle \geq 0$
 - Zero $\langle \mathbf{f}, \mathbf{f} \rangle = 0 \implies \mathbf{f} = 0$
- 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

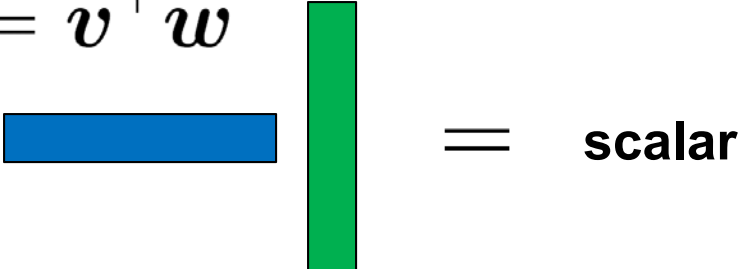
- Example of an inner product (just an example, inner product not required to be an integral)

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

Inner product of two functions is a number

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

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





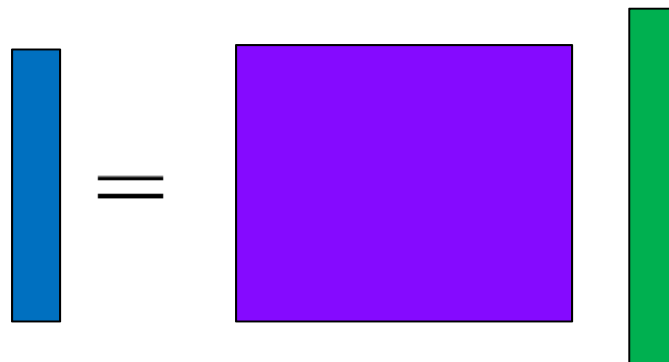
Linear Operators

- An operator \mathbf{C} maps a function \mathbf{f} in one Hilbert Space to another function \mathbf{g} in the same or another Hilbert Space.

- Linear Operator: $\mathbf{g} = \mathbf{C}\mathbf{f}$

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

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





Adjoins (Transposes)

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

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

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

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



Hilbert Space Outer Product

$f \otimes g$ is implicitly defined such that

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

Outer Product of two functions is an operator

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

$$\begin{aligned} v \otimes w &= vw^\top \\ vw^\top(z) &= \langle w, z \rangle v \end{aligned}$$

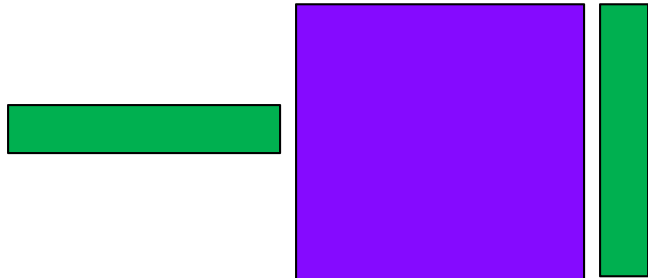
Reproducing Kernel Hilbert Space



- 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 $\mathbf{K}(x, y)$ is a function of two variables, such that:

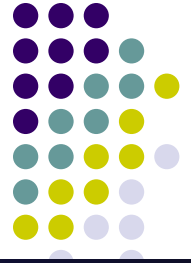
$$\int \int \mathbf{K}(x, y) \mathbf{f}(x) \mathbf{f}(y) dx dy > 0 \quad \forall \mathbf{f}$$

- This is a generalization of a positive definite matrix:

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


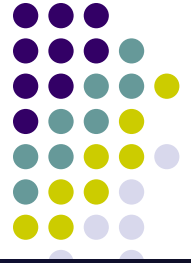
The diagram shows a horizontal green bar representing a row vector, followed by a large purple square representing a matrix, and a vertical green bar representing a column vector. These three elements are connected by a thin vertical line, and the entire expression is followed by a greater-than sign and a zero.

Gaussian Kernel



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

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



The Feature Function

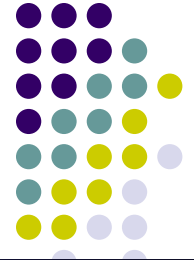
- Consider holding one element of the kernel fixed. We get a function of one variable which we call the feature function. The collection of feature functions is called the **feature map**.

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

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

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

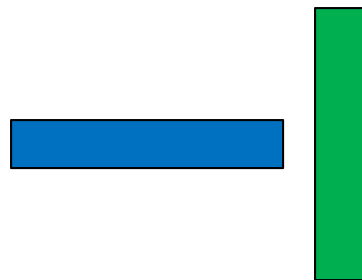
$$\phi_{1.5}(y) = \exp\left(\frac{\|1.5 - y\|_2^2}{\sigma^2}\right)$$



Defining the Inner Product

- Define the Inner Product as:

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



= scalar

- Note that:

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

Reproducing Kernel Hilbert Space

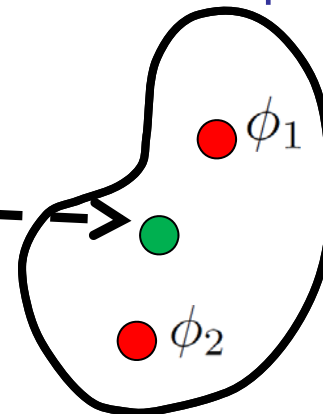


- Consider the set of functions that can be formed with linear combinations of these feature functions:

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

- We define the Reproducing Kernel Hilbert Space \mathcal{F} to be 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

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



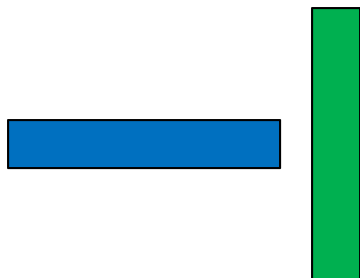


Reproducing Property

- It can now be derived that the inner product of a function \mathbf{f} with ϕ_X , evaluates a function at point \mathbf{x} :

$$\begin{aligned}\langle \mathbf{f}, \phi_x \rangle &= \left\langle \sum_j \alpha_j \phi_{x_j}, \phi_x \right\rangle \\ &= \sum_j \alpha_j \langle \phi_{x_j}, \phi_x \rangle && \text{Linearity of inner product} \\ &= \sum_j \alpha_j \mathbf{K}(x_j, x) && \text{Definition of kernel} \\ &= \mathbf{f}(x)\end{aligned}$$

Remember that
 $\mathbf{K}(x_j, x) := \phi_{x_j}(x)$



= **scalar**

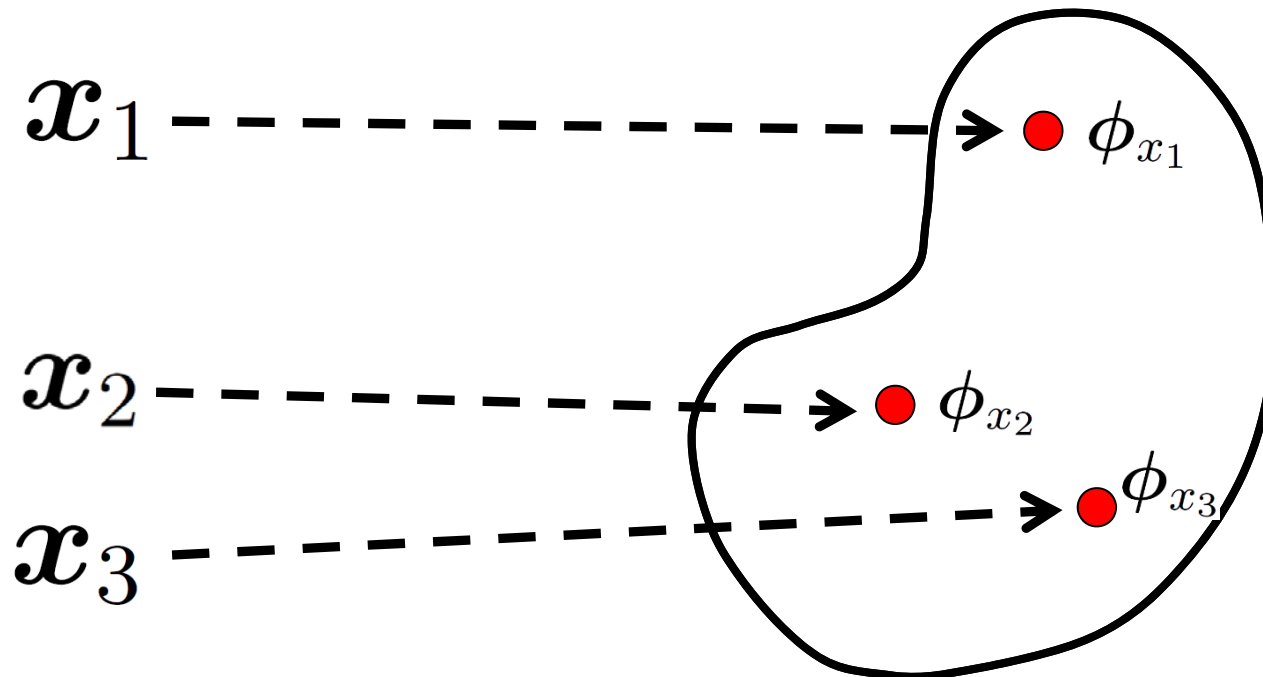


SVM Kernel Intuition

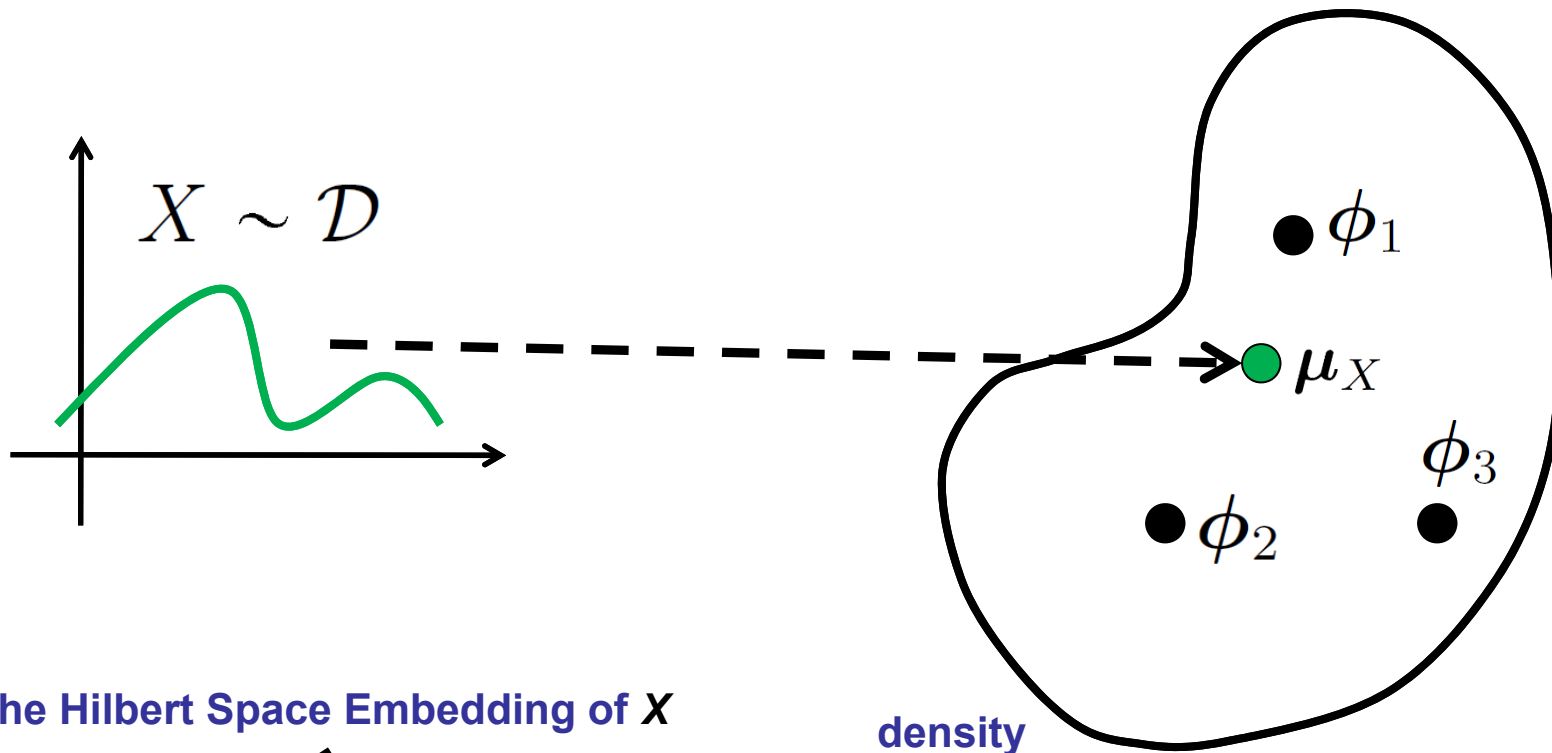
$$\min_{w,b} \frac{1}{2} \mathbf{w}^\top \mathbf{w} + C \sum_j \xi_j$$

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

Maps data points to RKHS Feature Functions!

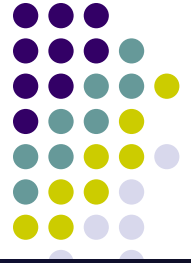


How To Embed Distributions (Mean Map) [Smola et al. 2007]



The Hilbert Space Embedding of X

$$\mu_X(\cdot) = \mathbb{E}_{X \sim \mathcal{D}}[\phi_X] = \int \text{density} \, p_{\mathcal{D}}(X) \phi_X(\cdot) dX$$



Mean Map cont.

- Mean Map $\mu_X = \mathbb{E}_X[\phi_X]$
- 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)

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

Data point



Example (Discrete)

- Consider a random variable X that takes the values $1, 2, 3, 4$. We want to embed it into an **RKHS**. Which **RKHS**?
- The RKHS of 4 dimensional vectors in \mathbf{R}^4 . The feature functions in this RKHS are:

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

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

$$\mu_X = \begin{pmatrix} \mathbb{P}[X = 1] \\ \mathbb{P}[X = 2] \\ \mathbb{P}[X = 3] \\ \mathbb{P}[X = 4] \end{pmatrix} \quad \text{Embedding equal to marginal probability vector in the discrete case}$$

Mean Map cont.



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

- Why?

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

Embedding Joint Distribution of 2 Variables

[Smola et al. 2007]

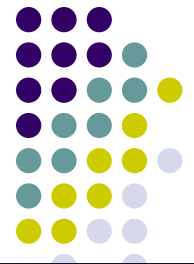


- Define the uncentered cross-covariance operator \mathcal{C}_{YX} implicitly such that

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

- Note now \mathbf{f} is in one Hilbert Space, while \mathbf{g} is in another.
- \mathcal{C}_{YX} will be our embedding of the joint distribution of \mathbf{X} and \mathbf{Y} .
- Note now \mathcal{C}_{YX} is an operator, just like $\mathbf{P}[\mathbf{X}, \mathbf{Y}]$ 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 \mathbf{X} and \mathbf{Y} :

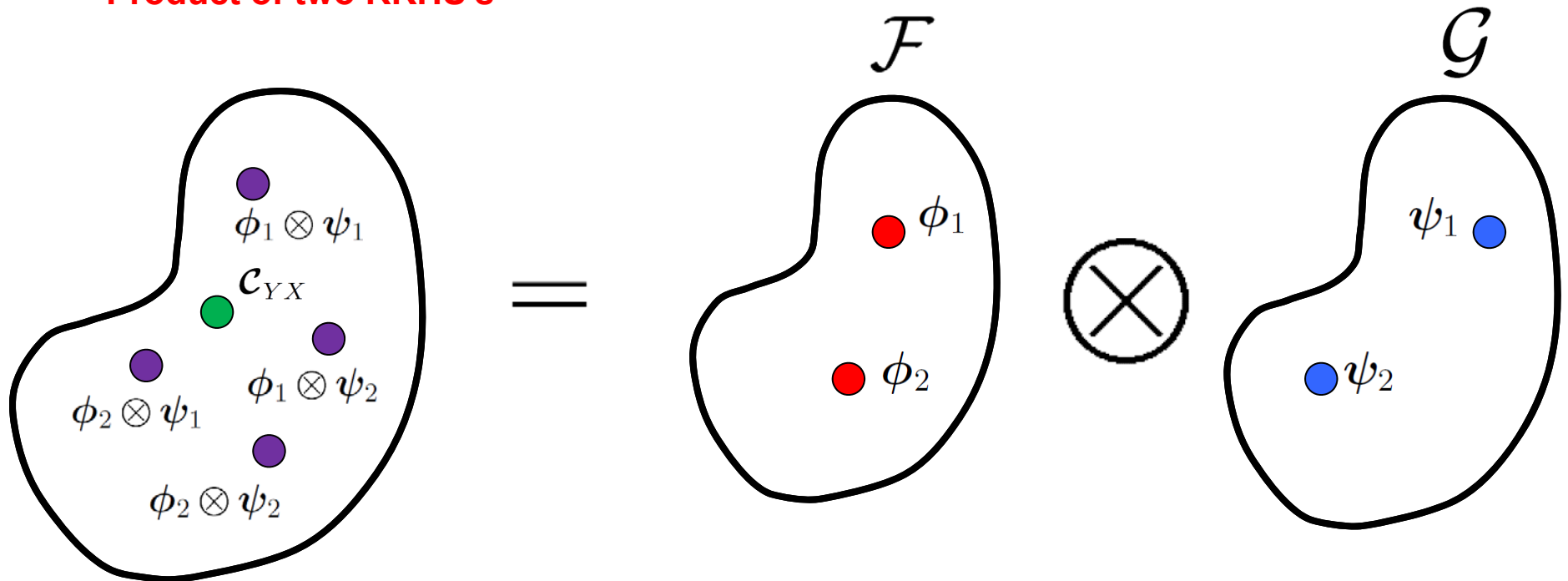
$$\text{Cov}(X, Y) = \mathbb{E}_{YX} [Y X]$$

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:

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

- If “outer product” is defined as:

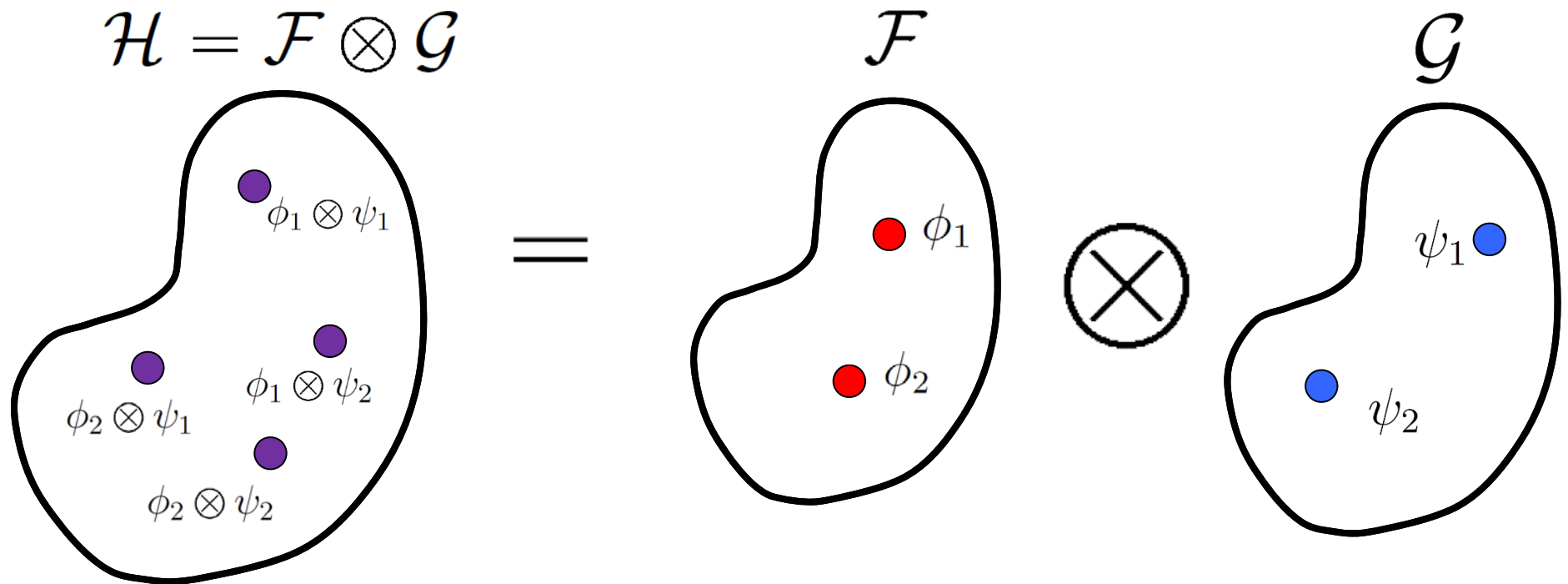
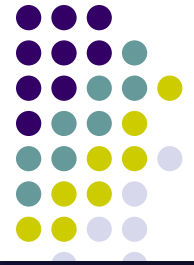
$$a \otimes b = (a, b)$$

- Then tensor product is:

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

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

Tensor Product of Two Vector Spaces



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

Cross Covariance Operator cont.



- **Proof:**

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



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:

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

- Intuition:

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



Conditional Embedding Cont.

- Conditional Embedding Operator:

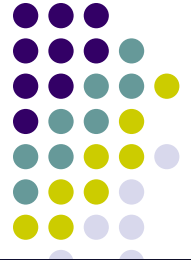
$$\mathbf{C}_{Y|X} = \mathbf{C}_{YX} \mathbf{C}_{XX}^{-1}$$

- Has Following Property:

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

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

$$\mathbb{P}[Y|X = 1] = \mathbb{P}[Y|X] \delta_1$$

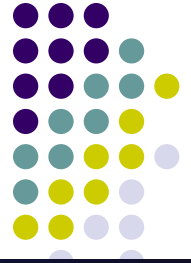


Why We Care

- 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:

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

Chain Rule:

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

Sum Rule in RKHS:

$$\mu_X = \mathcal{C}_{X|Y}\mu_Y$$

Chain Rule in RKHS:

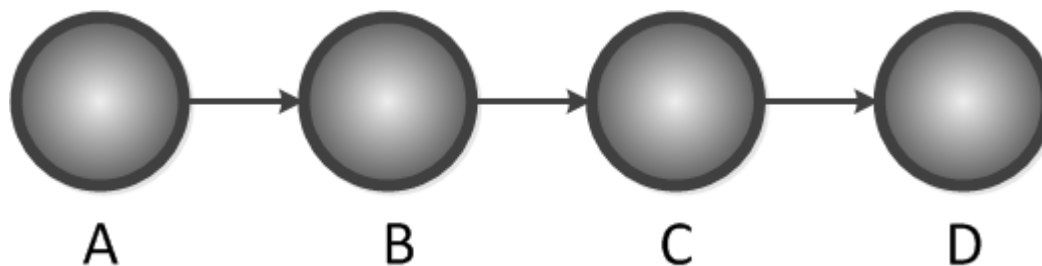
$$\mathcal{C}_{YX} = \mathcal{C}_{Y|X}\mathcal{C}_{XX} = \mathcal{C}_{X|Y}\mathcal{C}_{YY}$$

Kernel Graphical Models [Song et al. 2010,

Song et al. 2011]



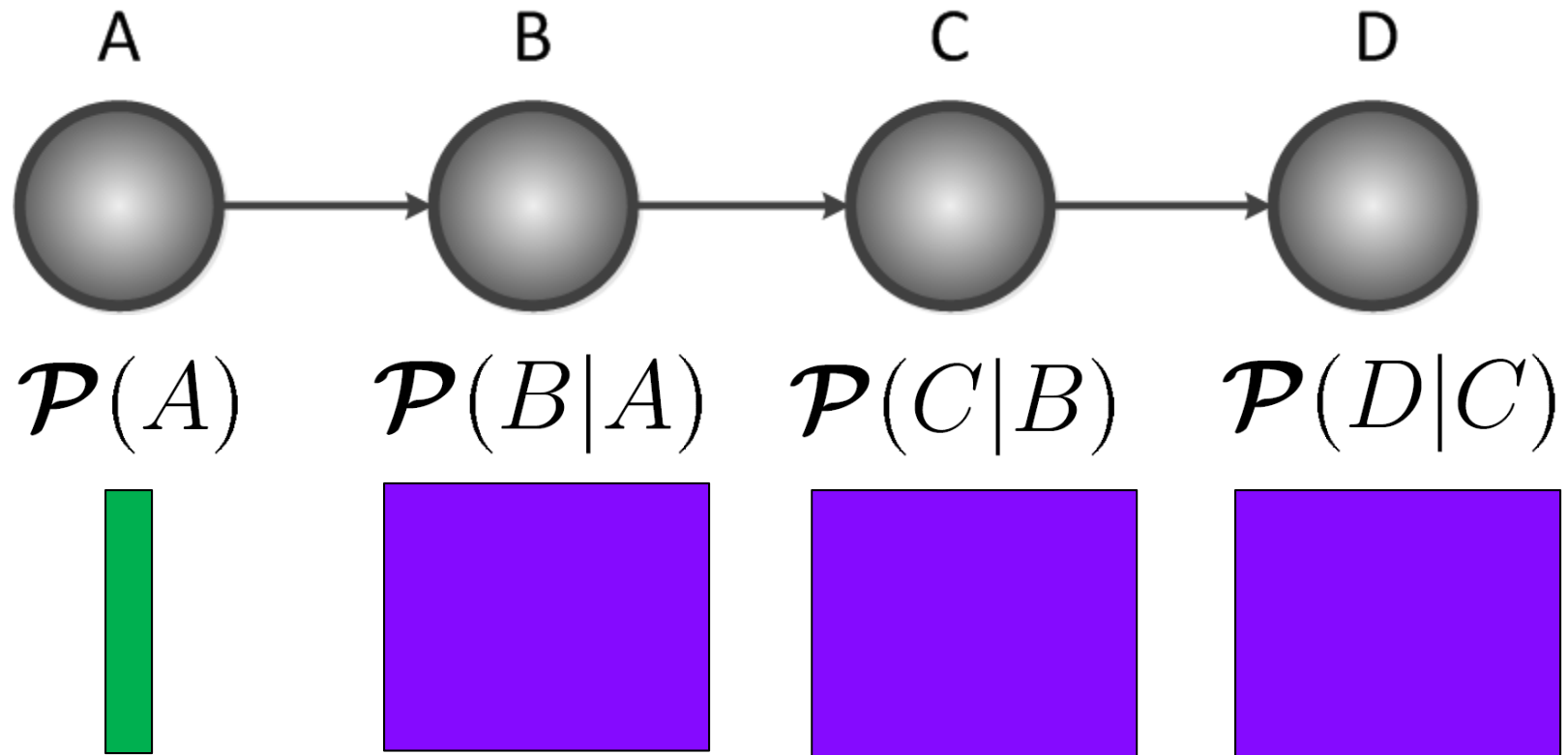
- The idea is to replace the CPTs with RKHS operators/functions.
- Let's do this for a simple example first.



- **We would like to compute** $\mathbb{P}[A = a, D = d]$

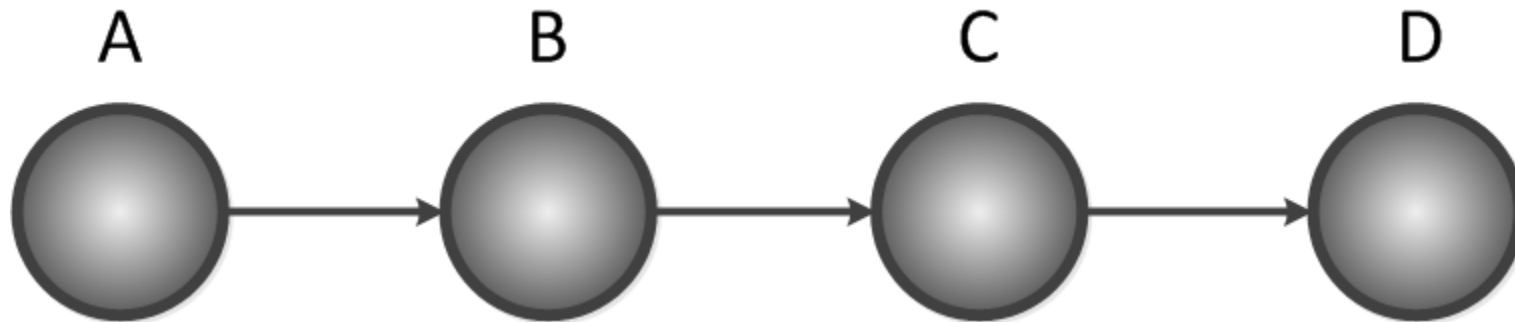


Consider the Discrete Case

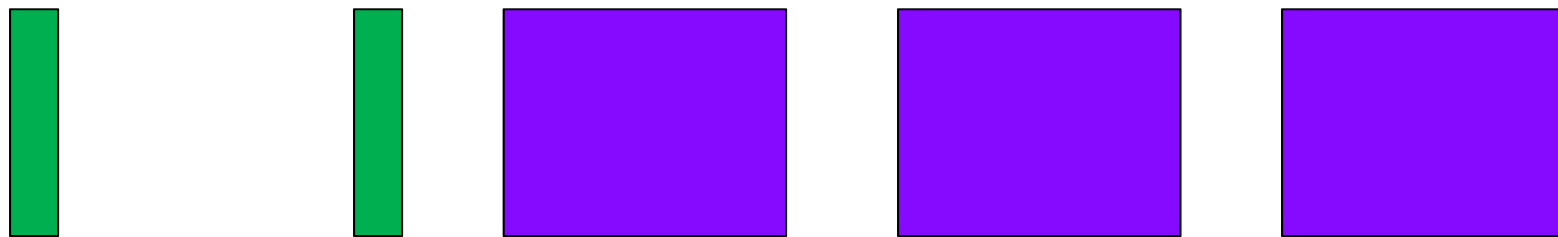




Inference as Matrix Multiplication



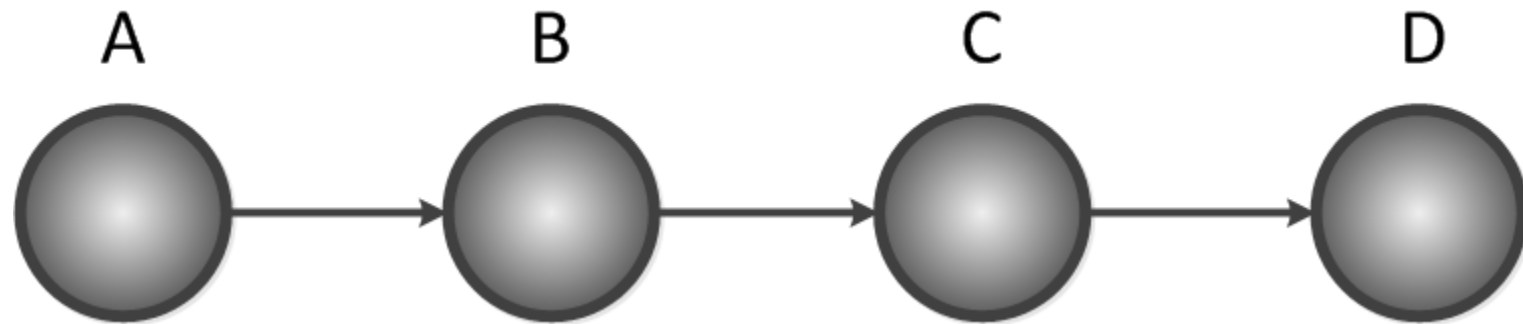
$$\mathcal{P}(D) = \mathcal{P}(A)\mathcal{P}(B|A)^\top \mathcal{P}(C|B)^\top \mathcal{P}(D|C)^\top$$



Oops....we accidentally integrated out A



Put A on Diagonal Instead

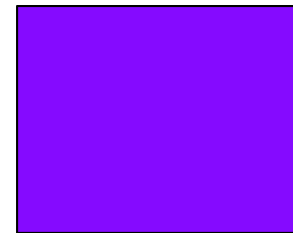
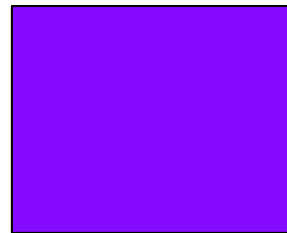
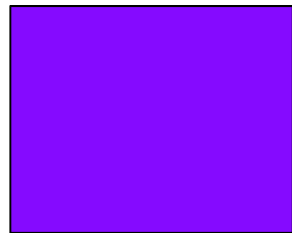
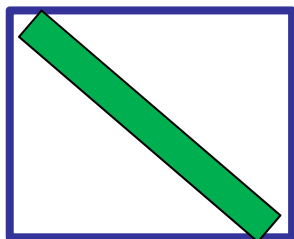


$$\mathcal{P}(\emptyset|A)$$

$$\mathcal{P}(B|A)$$

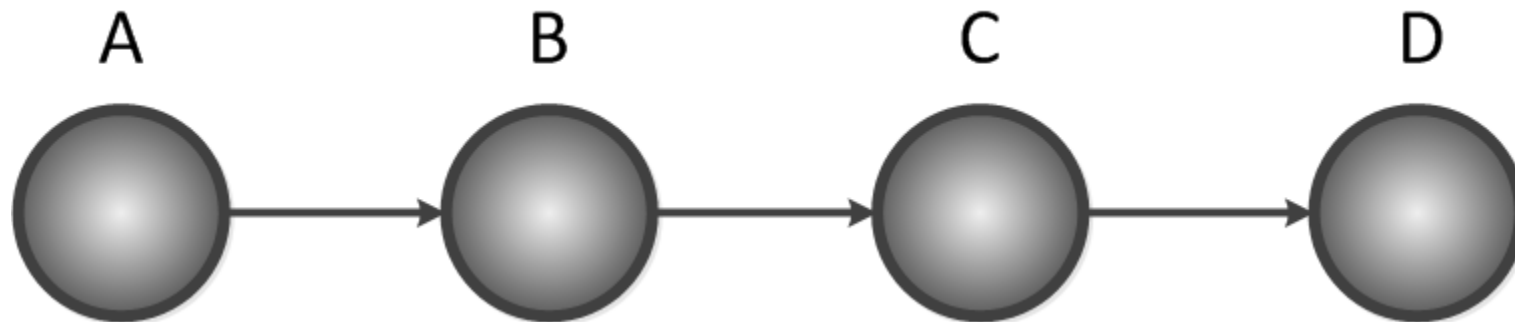
$$\mathcal{P}(C|B)$$

$$\mathcal{P}(D|C)$$

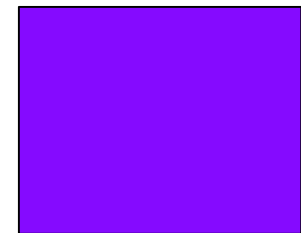
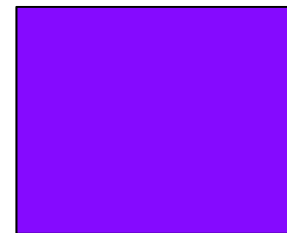
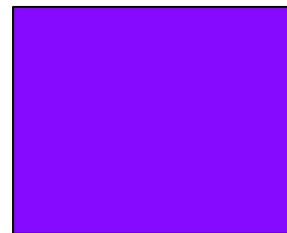
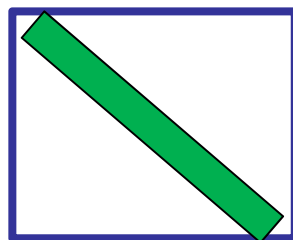
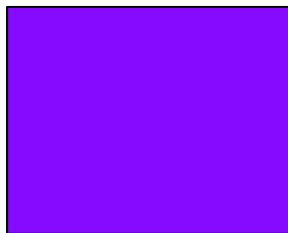




Now it works

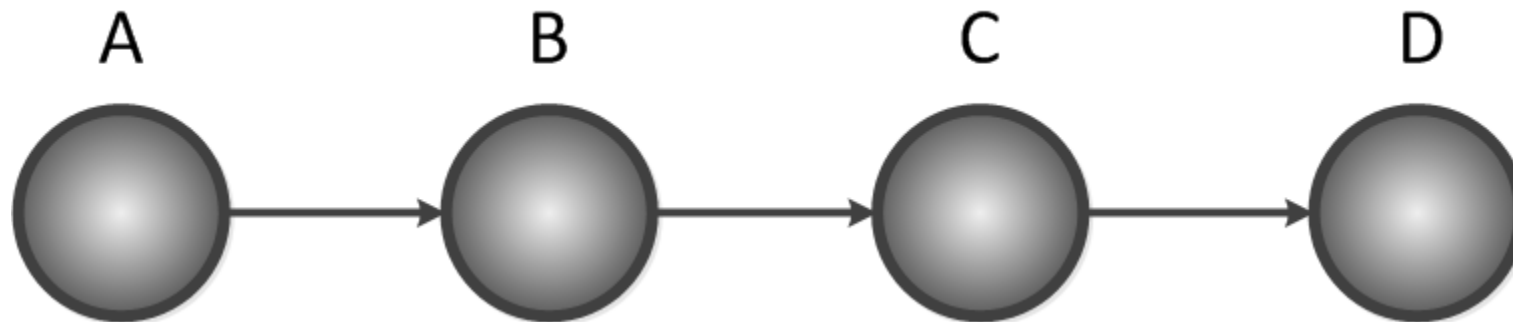


$$\mathcal{P}(A, D) = \mathcal{P}(\emptyset|A) \mathcal{P}(B|A)^\top \mathcal{P}(C|B)^\top \mathcal{P}(D|C)^\top$$





Introducing evidence

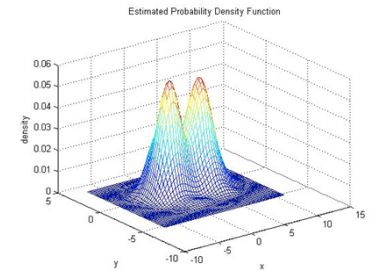
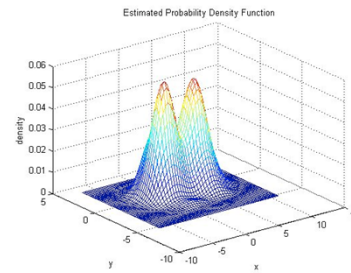
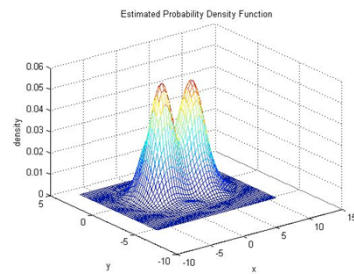
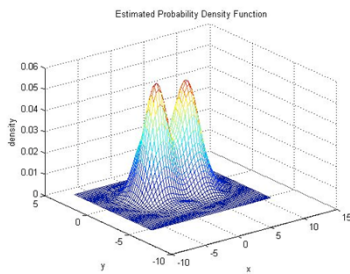
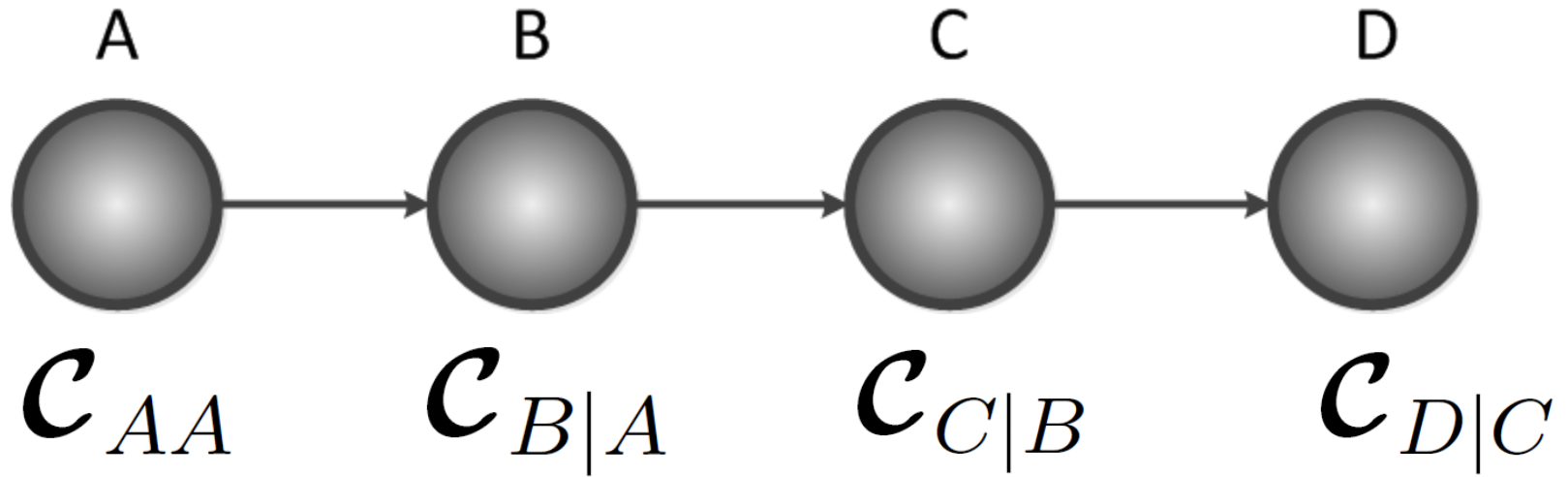


- Introduce evidence with delta vectors

$$\mathcal{P}(A = a, D = d) = \delta_a^\top \mathcal{P}(A, D) \delta_d$$

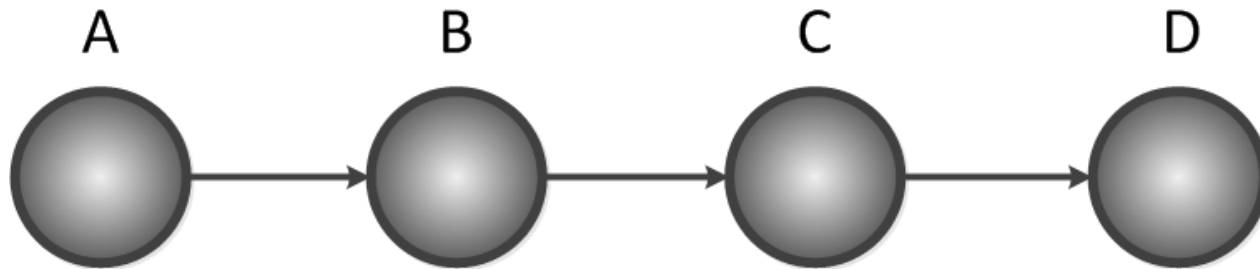


Now with Kernels





Sum-Product with Kernels

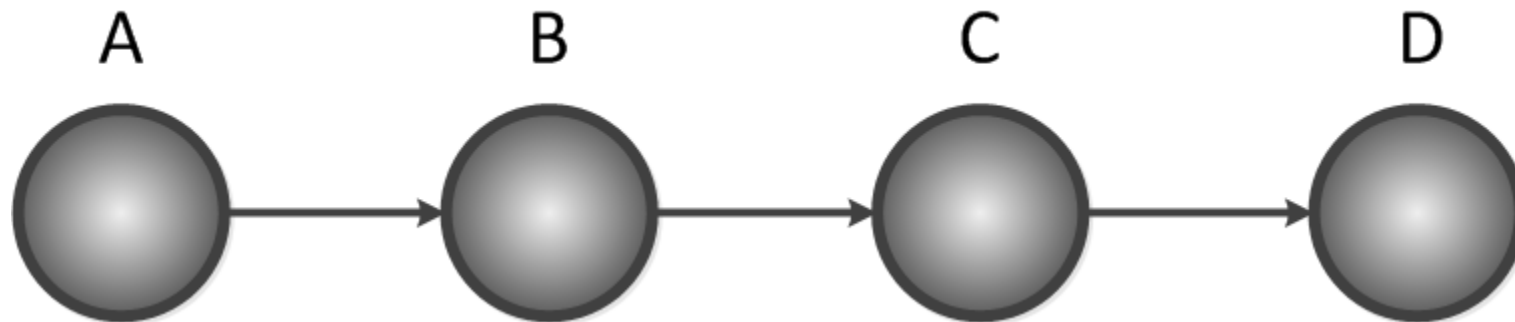


$$\mathcal{C}_{AB} = \mathcal{C}_{AA} \mathcal{C}_{B|A}^{\top}$$

$$\mathcal{C}_{AD} = \mathcal{C}_{AA} \mathcal{C}_{B|A}^{\top} \mathcal{C}_{B|C}^{\top} \mathcal{C}_{C|D}^{\top}$$



Sum-Product with Kernels



$$\text{some number} = \phi_a^\top \mathbf{C}_{A,D} \phi_d$$

Kernel Graphical Models: The Overall Picture



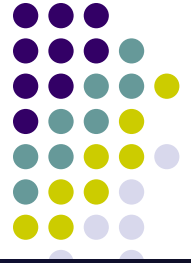
Naïve way to represent joint distribution of discrete variables is to store and manipulate a “huge” probability table.

Naïve way to represent joint distribution for many continuous variables is to use multivariate kernel density estimation.

Discrete Graphical Models allow us to factorize the “huge” joint distribution table into smaller factors.

Kernel Graphical Models allow us to factorize joint distributions of continuous variables into smaller factors.

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)

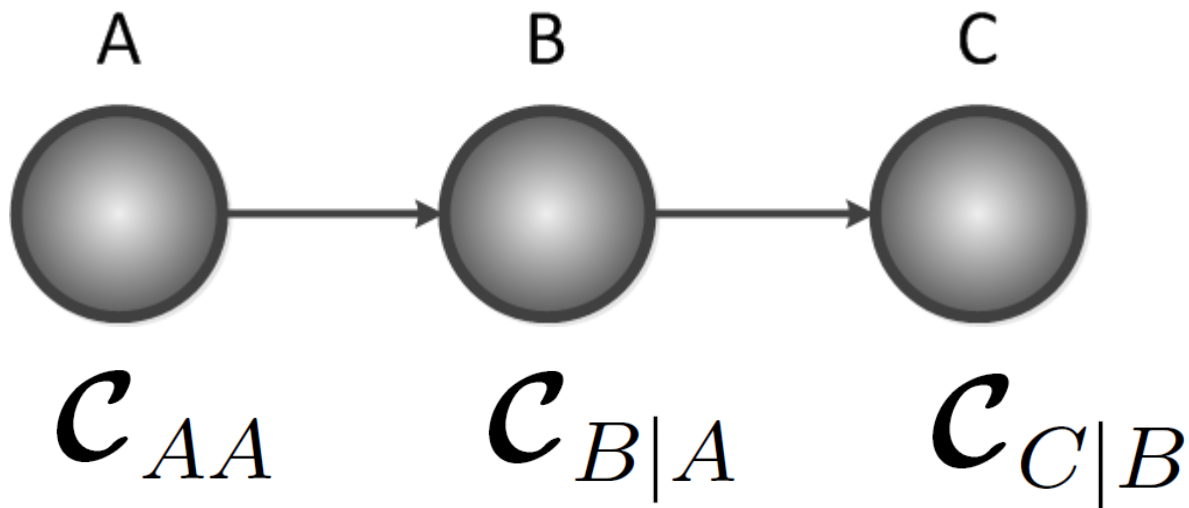


References

- 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.

Appendix:

Consider an Even Simpler Graphical Model



We are going to show how to estimate these operators from data.

The Kernel Matrix



$$\mathbf{K}_{XX} = \begin{bmatrix} \langle \phi_{x_1}, \phi_{x_1} \rangle & \cdots & \langle \phi_{x_1}, \phi_{x_N} \rangle \\ \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot \\ \langle \phi_{x_N}, \phi_{x_1} \rangle & \cdots & \langle \phi_{x_N}, \phi_{x_N} \rangle \end{bmatrix}$$

The matrix is annotated with a green double-headed arrow above it labeled N , indicating its width. A green arrow on the right points upwards, and another green arrow on the right points downwards, both labeled N , indicating its height.

Empirical Estimate Auto Covariance

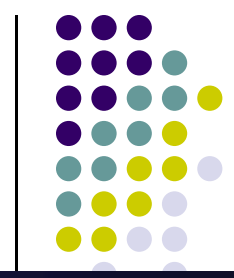


$$\mathbf{C}_{XX} = \mathbb{E}_X[\phi_X \otimes \phi_X]$$

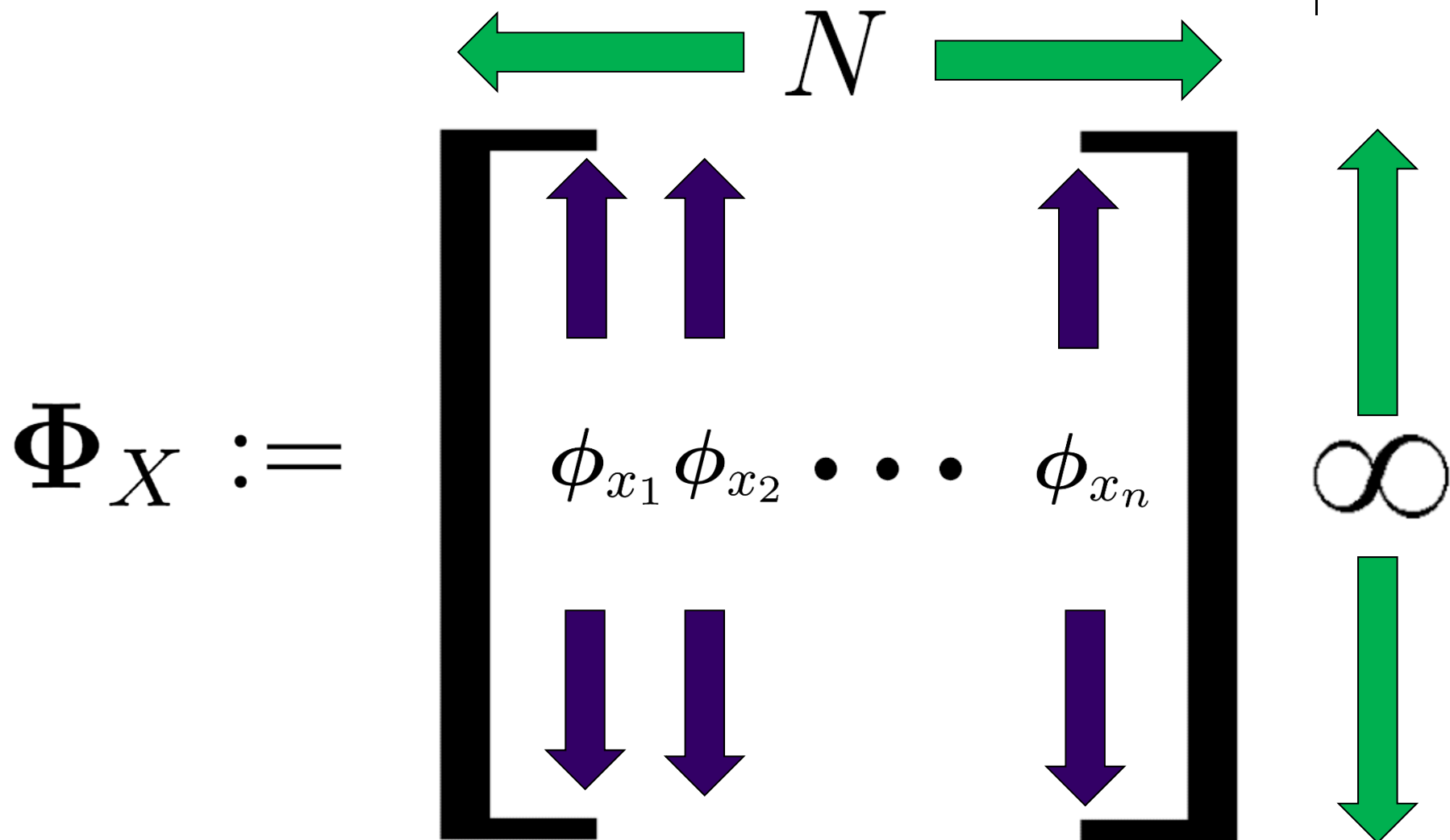
$$\hat{\mathbf{C}}_{XX} = \frac{1}{N} \sum_{n=1}^N \phi_{x_n} \otimes \phi_{x_n}$$

$$\hat{\mathbf{C}}_{XX} = \frac{1}{N} \Phi_X \Phi_X^\top$$

Defined on next slide



Conceptually,





Conceptually,

 Φ_X

$$\sum_{n=1}^N v_i \phi_{x_n} = \left[\begin{array}{ccc} \uparrow & \uparrow & \uparrow \\ \phi_{x_1} & \phi_{x_2} & \dots & \phi_{x_n} \\ \downarrow & \downarrow & \downarrow \end{array} \right] \begin{pmatrix} v_1 \\ v_2 \\ \dots \\ v_N \end{pmatrix}$$



Conceptually,

$$\begin{pmatrix} \phi_{x_1}^\top f \\ \phi_{x_2}^\top f \\ \dots \\ \phi_{x_n}^\top f \end{pmatrix} = \left[\begin{array}{ccc} \leftarrow & \phi_{x_1} & \rightarrow \\ \leftarrow & \phi_{x_2} & \rightarrow \\ \dots & \dots & \dots \\ \leftarrow & \phi_{x_n} & \rightarrow \end{array} \right] \Phi_X^\top f$$

The diagram illustrates the conceptual relationship between the matrix product $\Phi_X^\top f$ and the vector of inner products $(\phi_{x_1}^\top f, \phi_{x_2}^\top f, \dots, \phi_{x_n}^\top f)^\top$. The matrix Φ_X^\top is shown as a large bracketed structure with rows corresponding to the input features $\phi_{x_1}, \phi_{x_2}, \dots, \phi_{x_n}$. Purple arrows point from each row of Φ_X^\top to the corresponding element in the vector of inner products, indicating that the matrix multiplication is equivalent to computing these inner products.



Rigorously,

Φ_X is an operator that maps vectors in \mathbb{R}^N to functions in \mathcal{F}

such that:

$$\sum_{n=1}^N v_n \phi_{x_n} = \Phi_X v$$

Its adjoint (transpose) Φ_X^\top can then be derived to be:

$$\begin{pmatrix} \langle \phi_{x_1}, f \rangle \\ \langle \phi_{x_2}^\top, f \rangle \\ \dots \\ \langle \phi_{x_n}^\top, f \rangle \end{pmatrix} = \Phi_X^\top f$$

Empirical Estimate Cross Covariance



$$\mathbf{C}_{YX} = \mathbb{E}[\phi_Y \otimes \phi_X]$$

$$\hat{\mathbf{C}}_{YX} = \frac{1}{N} \sum_{n=1}^N \phi_{y_n} \otimes \phi_{x_n}$$

$$\hat{\mathbf{C}}_{YX} = \frac{1}{N} \mathbf{\Phi}_Y \mathbf{\Phi}_X^\top$$



Getting the Kernel Matrix

- It can then be shown that,

$$\mathbf{\Phi}_X^\top \mathbf{\Phi}_X = \mathbf{K}_{XX} \quad \mathbf{K}_{XX}(i, j) := \langle \phi_{x_i}, \phi_{x_j} \rangle$$

- This is finite and easy to compute!! 😊
- However, note that the estimates of the covariance operators are **not** finite since:

$$\hat{\mathbf{C}}_{XX} = \frac{1}{N} \mathbf{\Phi}_X \mathbf{\Phi}_X^\top$$