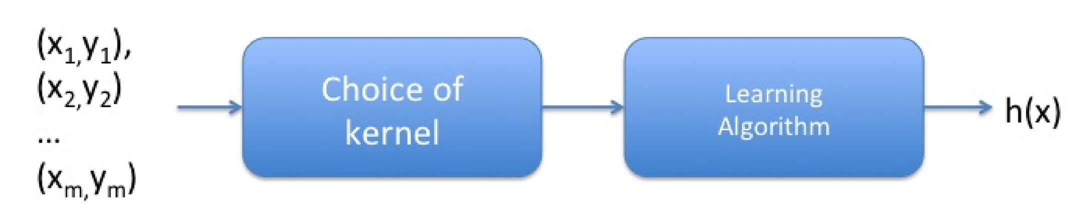
Simultaneous Twin Kernel Learning using Polynomial Kernel Transfomations for Structured Prediction

Chetan Tonde Ahmed Elgammal
{cjtonde,elgammal}@cs.rutgers.edu
Department of Computer Science, Rutgers, The State University of New Jersey.

IEEE 2014 Conference on Computer Vision and Pattern Recognition

Objective



How to choose the right kernel functions $k(\cdot,\cdot)$ and $g(\cdot,\cdot)$ so as to perform better *structured prediction*?

Introduction

- In *structured prediction* we learn a prediction function $f : \mathcal{X} \to \mathcal{Y}$ from an input domain $\mathcal{X}$ to an output domain $\mathcal{Y}$.
- We formulate an *auxiliary evaluation function* $h : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$, such that,

$$y^* = f(x) = \arg\max_{y \in \mathcal{Y}} h(x, y) \qquad (1)$$

- Kernel methods [3] define kernel maps $k(x,\cdot) : \mathcal{X} \to \mathcal{K}$ and $g(y,\cdot) : \mathcal{Y} \to \mathcal{G}$ *jointly on input and outputs*.
- Structured data is high dimensional and highly structured, and the choice of kernel functions is difficult so can we *learn both input and output kernel functions simultaneously*.
- Twin Gaussian Processes [2] as an example model input/output using Gaussian Process prior with *covariance functions represented by kernel matrices* $\mathbf{K}$ and $\mathbf{G}$.

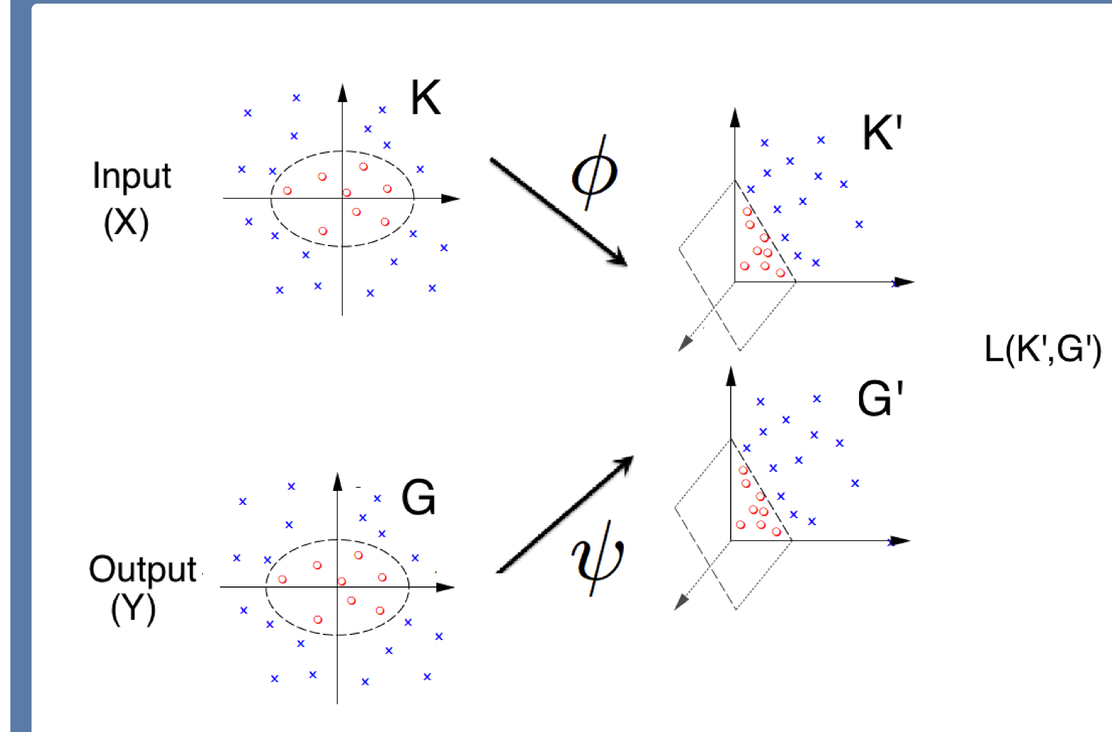
Polynomial Kernel Transformation

- **Theorem** [FitzGerald *et al.* (1995) [1]]: If there exists a continuous function $\phi : \mathbb{R} \to \mathbb{R}$, such that, $[\mathbf{K}']_{i,j} = \phi([\mathbf{K}]_{i,j})$ then, $\mathbf{K}'$ is positive definite for any SPD matrix $\mathbf{K}$, if and only if, $\phi(\cdot)$ it is real entire and of the form below,

$$\phi(t) = \sum_{i=0}^{\infty} \alpha_i t^i, \text{ with } \alpha_i \geq 0 \text{ for all } i \geq 0. \qquad (2)$$

- *Example*: The exponential function $\phi(t) = e^t$, $\phi(t) = e^t = 1 + \frac{t}{1} + \frac{t^2}{2!} + \frac{t^3}{3!} + \ldots$, with $\alpha_i = \frac{1}{i!}$

Twin Kernel Transfomations



Algorithm

- The *statistical dependence* between two feature spaces $\mathcal{X}$ and $\mathcal{Y}$ is given by, $HSIC(P(\mathbf{x},\mathbf{y}),\mathcal{K},\mathcal{G}) := ||\mathbf{M}_{\mathbf{x},\mathbf{y}}||^2_{HS}$ where, $\mathbf{M}_{\mathbf{x},\mathbf{y}} := \mathbf{E}_{\mathbf{x},\mathbf{y} \in P(\mathbf{x},\mathbf{y})}[(k(\mathbf{x},\cdot) - \mu_{\mathbf{x}}) \otimes (g(\mathbf{y},\cdot) - \mu_{\mathbf{y}})]$
- Empirically, $\overline{HSIC}(X \times Y, \mathcal{K}, \mathcal{G}) = (m-1)^{-2} trace(\mathbf{HKHGH})$ where $[\mathbf{H}]_{i,j} := \delta_{ij} - m^{-1}$
- After *approximating* and *adding regularization* (α_i, β_j) we get the final problem as,

$$\text{maximize} \sum_{i=0}^{d_1} \sum_{j=0}^{d_2} \alpha_i \beta_j \mathbf{C}_{i,j} \qquad (3)$$

$$\text{subject to, } ||\boldsymbol{\alpha}||_2 = 1, ||\boldsymbol{\beta}||_2 = 1, \boldsymbol{\alpha} \geq 0, \boldsymbol{\beta} \geq 0$$

where $[\mathbf{C}]_{i,j} = \overline{HSIC}(\mathbf{K}^{(i)}, \mathbf{G}^{(j)})$.

- **Theorem**: The solution $(\boldsymbol{\alpha}^*, \boldsymbol{\beta}^*)$ to the above optimization problem is given by, the *first left and right singular vectors* of the $\mathbf{C}$-matrix.
- For $d_1 = d_2 = 1$, $\phi(t) = t$ and $\psi(t) = t$ corresponds to *no mapping*.

Modified Twin Gaussian Processes

- **TGP with KL-Divergence**

$$\mathbf{y}^* = \arg\min_{y} D_{KL}((\psi(\boldsymbol{G}_{\mathbf{Y}\cup y})||\phi(\boldsymbol{K}_{\mathbf{X}\cup x}))$$

- **TGP with $\overline{HSIC}$**:

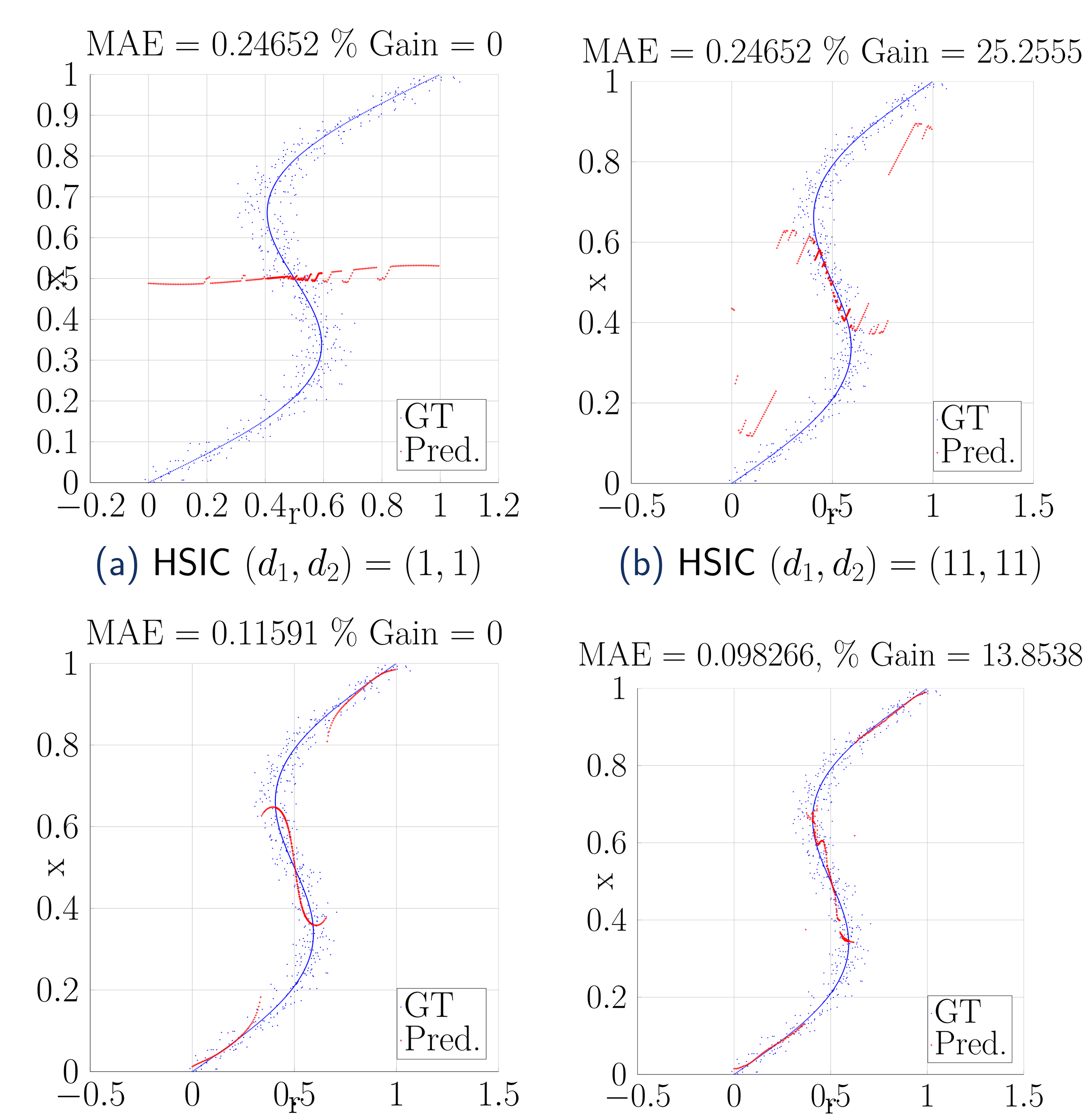
$$\mathbf{y}^* = \arg\max_{y} \overline{HSIC}((\psi(\boldsymbol{G}_{\mathbf{Y}\cup y}), \phi(\boldsymbol{K}_{\mathbf{X}\cup x}))$$

Experiments

- Empirical performance measure:

$$\%\ Gain = \left(1 - \frac{Error_{(mapping)}}{Error_{(no\ mapping)}}\right) \times 100$$

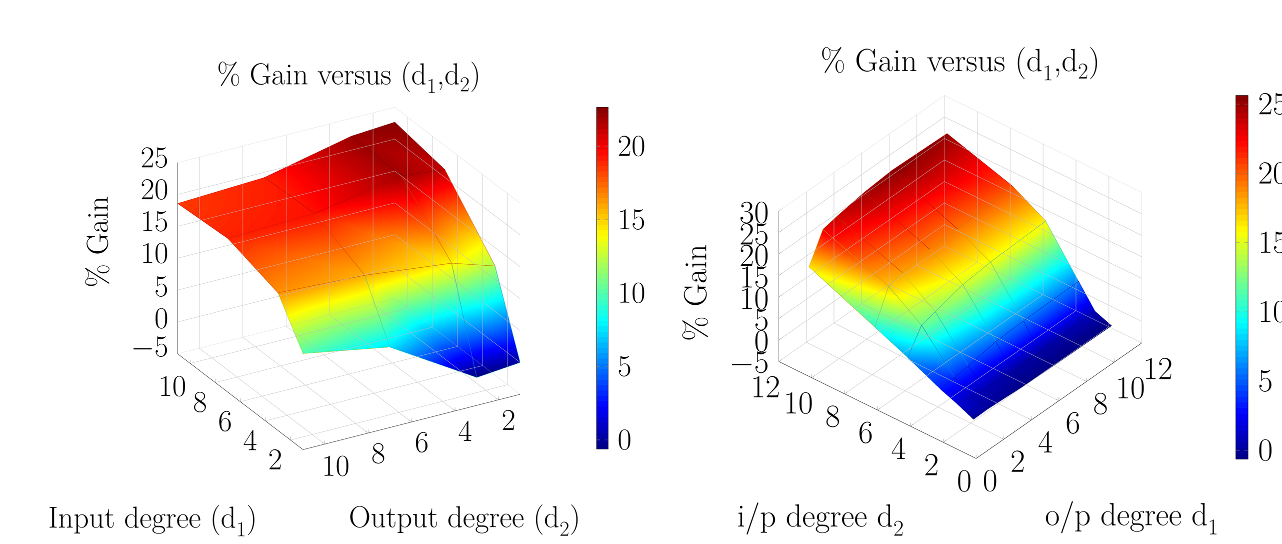
- We use RBF kernels on the input and output, $k(\mathbf{x}_i, \mathbf{x}_j) = e^{(-\gamma_{\mathbf{x}}||x_i - \mathbf{x}_j||^2)}$, $g(\mathbf{y}_i, \mathbf{y}_j) = e^{(-\gamma_{\mathbf{y}}||y_i - \mathbf{y}_j||^2)}$
- **S-Shape regression**: 1d input-output problem, $r \in (0, 1)$, $r = x + 0.3sin(2x\pi) + \epsilon$, with $\epsilon \sim \mathcal{N}(0, 0.05)$, multivalued, discontinuous, noisy.



(a) HSIC $(d_1, d_2) = (1, 1)$ (b) HSIC $(d_1, d_2) = (11, 11)$

(c) KL-Div $(d_1, d_2) = (1, 1)$ (d) KL-Div $(d_1, d_2) = (1, 11)$

Figure: Regression on S-Shape dataset with HOTGP (Fig.1c-1d) and HOHSIC (Fig.1a-1b)



(a) *% Gain* in Mean Abs. Error for KL-Div. criterion. (b) *% Gain* in Mean Abs. Error for HSIC criterion.

- **USPS Handwritten Digit Reconstruction**:

Approach	MAE	Approach	MAE
NN	0.341	KRR	0.250
SVR	0.250	KDE	0.260
$SOAR_{krr}$	0.233	$SOAR_{svr}$	0.230
HSIC (wo/map)	**0.3399**	KL-Div (wo/map)	**0.21508**
HSIC (w/map)	**0.3327**	KL-Div (w/map)	**0.21084**
% Gain	**2.4842 %**	*% Gain*	**1.9924 %**

Table: *% Gain* over for the KL-Div. and HSIC criterion.

- **Poser**: Synthetic human motion sequences; Input: shape descriptor, Output: x, y, z joint angles.

Criterion - (d_1, d_2)	*% Gain*
KL-Divergence - $(1, 11)$	**6.39 %**
HSIC - $(11, 11)$	**1.2613 %**

Table: *%Gain* w/ and w/o mapping for KL-Div. and HSIC.

- **Human Eva-I** : Human motion sequences; Input: HoG features Output: x, y, z joint positions.

Features	Crit.	wo/map	w/ map	Gain %
HoG (C1C2C3)	KL-Div	45.1729	42.8783	**5.0796 %**
	HSIC	171.4085	171.3766	**0.018613 %**
HoG (C1)	KL-Div	34.2885	33.4262	**2.5147 %**
	HSIC	171.4085	171.3769	**0.018427 %**
HoG (C2)	KL-Div	31.9928	31.5792	**1.2928 %**
	HSIC	171.4085	171.3755	**0.019237 %**
HoG (C3)	KL-Div	30.9279	30.4928	**1.4067 %**
	HSIC	171.4085	171.3762	**0.018835 %**

Table: *%Gain* for KL-Div. $(1, 11)$ and HSIC $(11, 11)$

Conclusion

We propose a novel, efficient and effective method for learning the kernels using polynomial kernel transformations for structured prediction problems.

References

[1] C. H. FitzGerald, et al. *Functions that preserve families of positive semidefinite matrices. LAA, '95.*

[2] L. Bo and C. Sminchisescu. *Twin Gaussian Processes for Structured Prediction. IJCV, 2010.*

[3] Nowozin, S. et al. *Structured Learning and Prediction in Computer Vision Now Pub. Inc, 2011*