

# Information Geometry of Diffusion Kernels

---

Koji Tsuda

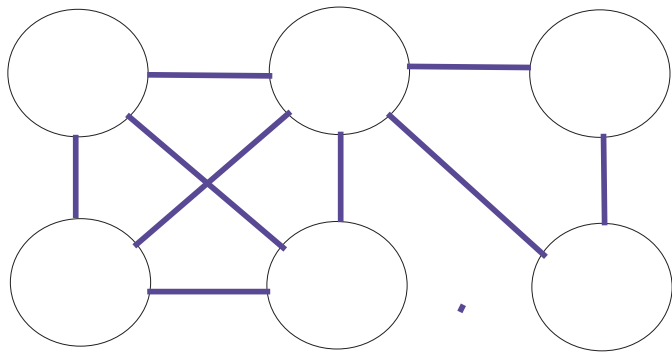
Max Planck Institute for Biological Cybernetics  
AIST Computational Biology Research Center

---

# Kernels Among Graph Nodes

---

- Input: Undirected Graph with  $n$ -nodes
- Output:  $n \times n$  Positive Definite Kernel Matrix  $K$
- How to design this mapping?



$$\begin{pmatrix} 1.00 & 0.46 & 0.46 & 0.18 & 0.02 & 0.00 \\ & 1.00 & 0.18 & 0.39 & 0.07 & 0.02 \\ & & 1.00 & 0.39 & 0.07 & 0.02 \\ & & & 1.00 & 0.37 & 0.12 \\ & & & & 1.00 & 0.60 \\ & & & & & 1.00 \end{pmatrix}$$

## Diffusion Kernel (Kondor and Lafferty, ICML 2002)

---

- $A$ : Graph adjacency matrix,  $D$ : Diagonal matrix of degrees
- $L = D - A$ : Graph Laplacian matrix
- Diffusion kernel

$$K = \frac{1}{Z(\beta)} \exp(-\beta L),$$

where  $\beta > 0$ : degree of diffusion, and  $Z(\beta) = \text{tr}(\exp(-\beta L))$ .

- Understood via physical intuitions (e.g. random walks)

# Goal of this talk

---

- Analyze the diffusion kernel in light of *Information Geometry*
- Better understanding
- Further Extensions
  - Locally Constrained Diffusion Kernels (Tsuda and Noble, ISMB 2004)

# Von Neumann entropy & Divergence

---

- Von Neumann entropy

$$E(K) = -\text{tr}[K \log K], \quad K \succ 0, \quad \text{tr}(K) = 1$$

- $\lambda_i$ :  $i$ -th eigenvalue of  $K$  ( $\sum_i \lambda_i = 1$ )
- Von Neumann Entropy is the Shannon entropy of eigenvalues

$$-\text{tr}(K \log K) = -\sum_i \lambda_i \log \lambda_i$$

- Von Neumann Divergence (Quantum Relative Entropy)

$$D(K, K_0) = \text{tr}[K \log K - K \log K_0], \quad K, K_0 \succ 0, \quad \text{tr}(K) = \text{tr}(K_0) = 1$$

# Diffusion Kernel by Maximum Entropy (Tsuda and Noble, ISMB2004)

---

- Maximizing entropy leads to the diffusion kernel  $\exp(-\beta L)$

$$\min_K \text{tr}(K \log K), \quad \text{tr}(K) = 1, \text{tr}(KL) \leq c$$

- $\mathbf{x}_1, \dots, \mathbf{x}_n$ : Embedded nodes in *feature space* ( $K_{ij} = \mathbf{x}_i^\top \mathbf{x}_j$ )
- $\text{tr}(KL)$ : Sum of Euclidean distances between connected nodes

$$\text{tr}(KL) = \sum_{i \sim j} \|\mathbf{x}_i - \mathbf{x}_j\|^2$$

- Projection of  $I/n$  in terms of the divergence

$$\min_K D(K, I/n), \quad \text{tr}(K) = 1, \text{tr}(KL) \leq c$$

# Information Geometry of PD matrices

---

- $e$ -flat subspace of matrices

$$\mathcal{E} = \{K \mid K = \exp\left(\sum_{i=1}^n \alpha_i \log(K_i)\right), \alpha_i \in \mathfrak{R}\}$$

- $m$ -flat subspace of matrices

$$\mathcal{M} = \{K \mid K = \sum_{i=1}^n \alpha_i K_i, \alpha_i \in \mathfrak{R}\}$$

or in implicit form

$$\mathcal{M} = \{K \mid \text{tr}(KL_i) = c_i, c_i \in \mathfrak{R}\}$$

# Projections

---

- $e$ -projection

$$\min_{K \in \mathcal{S}} D(K, K_0)$$

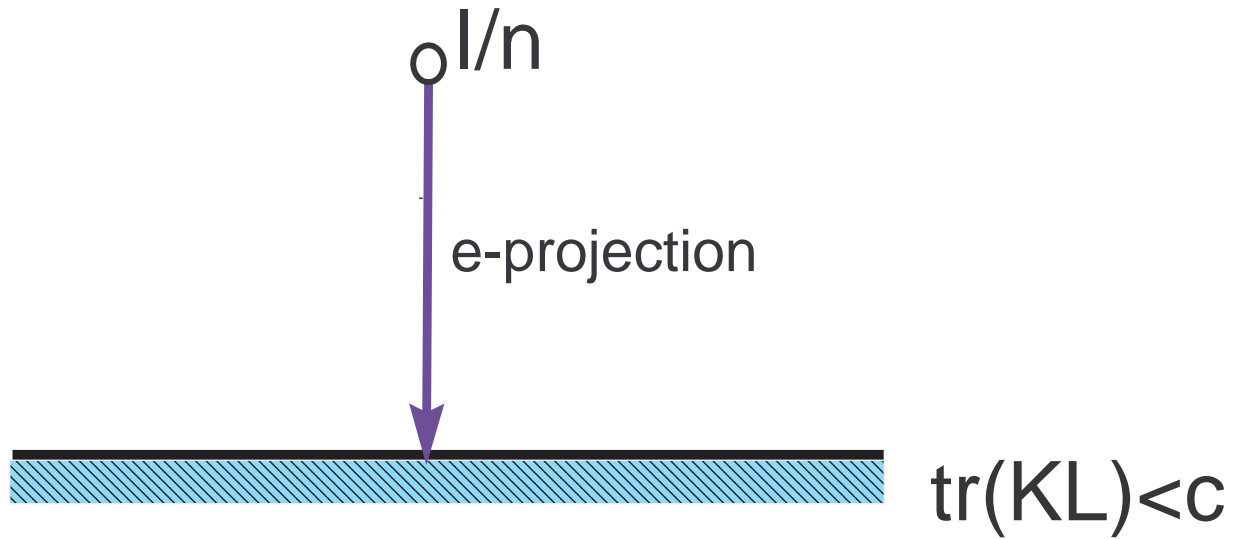
- $m$ -projection

$$\min_{K \in \mathcal{S}} D(K_0, K)$$

- $e$ -projection to  $m$ -flat subspace is unique
- $m$ -projection to  $e$ -flat subspace is unique
- Induced optimization problems are convex

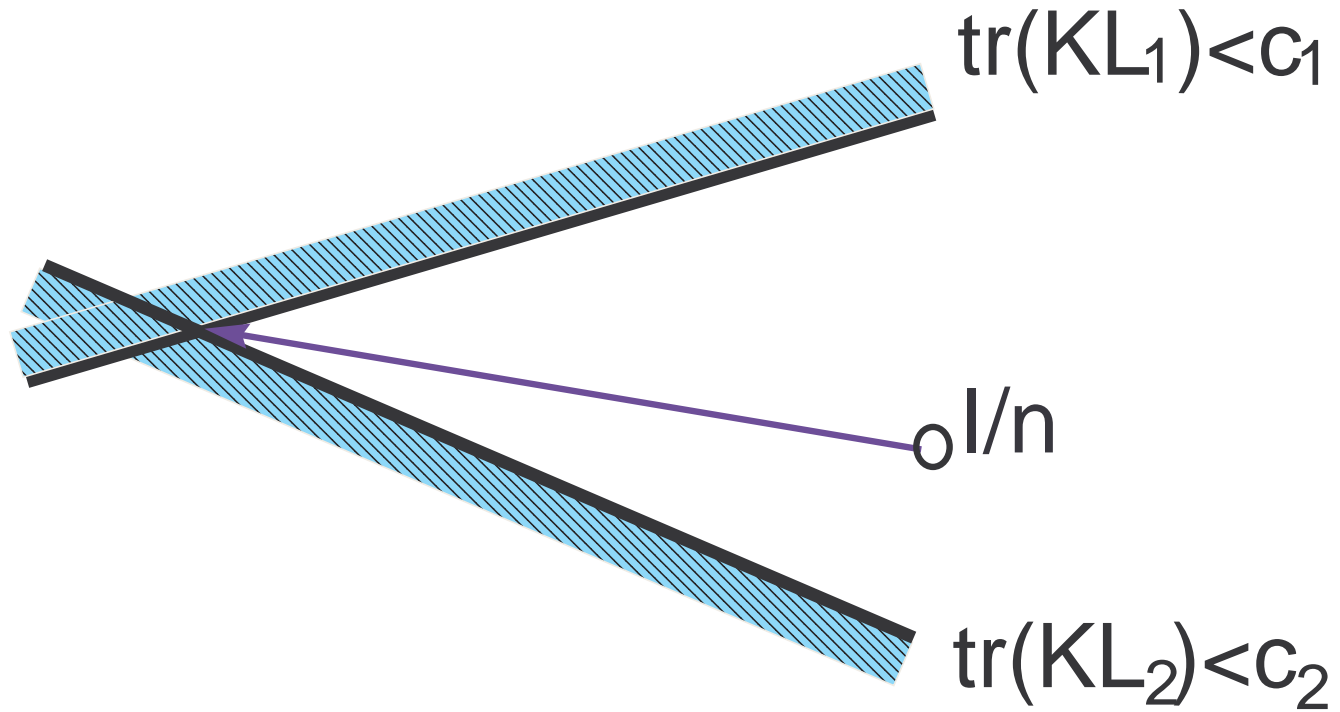
# Visualizing the max entropy theorem

---



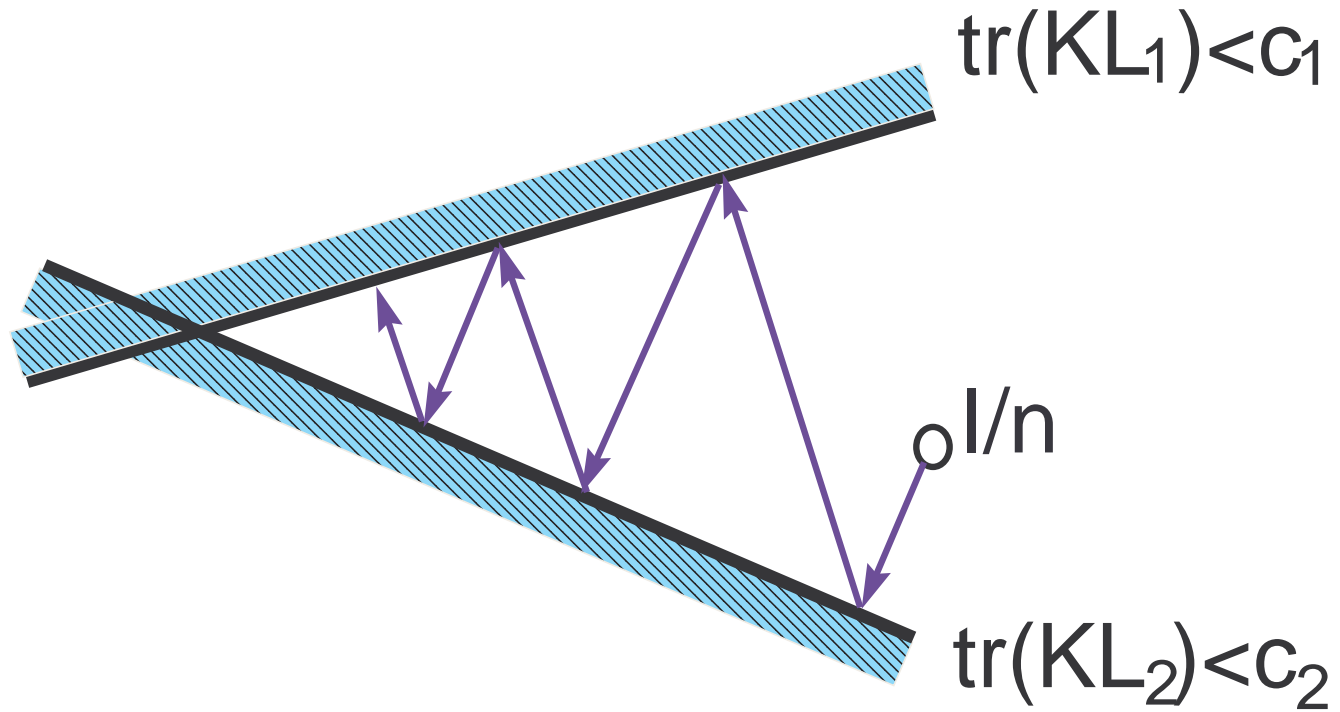
# Extention to multiple constraints

---



# Successive e-projections

---



# Successive e-projection

---

- Initial:  $K_0 = I/n$
- $t$ -th Iteration: Pick one constraint

$$\min_{K_t} D(K_t, K_{t-1}) \quad \text{tr}(K) = 1, \text{tr}(K L_t) \leq c_t$$

- Line search for one dual variable
- Approximate line search: *DefiniteBoost*
- See our NIPS paper (this year) for algorithms and convergence proof

# Locally Constrained Diffusion Kernel

---

- Since the diffusion kernel constrains the sum of distances, each distance has extremely high variance.
- $\{s_j, t_j\}_{j=1}^m$ : Node pairs connected by  $m$  edges
- Constrain each distance individually

$$\|\mathbf{x}_{s_j} - \mathbf{x}_{t_j}\|^2 \leq \gamma, \quad j = 1, \dots, m.$$

- Optimization Problem

$$\min_K \operatorname{tr}(K \log K), \quad \operatorname{tr}(K) = 1, \operatorname{tr}(K L_j) \leq \gamma, \quad j = 1, \dots, m,$$

where

$$[L_j]_{st} = \begin{cases} 1 & (s = s_j, t = s_j) \text{ or } (s = t_j, t = t_j) \\ -1 & (s = s_j, t = t_j) \text{ or } (s = t_j, t = s_j) \\ 0 & \text{otherwise} \end{cases}$$

# Protein Network in Yeast

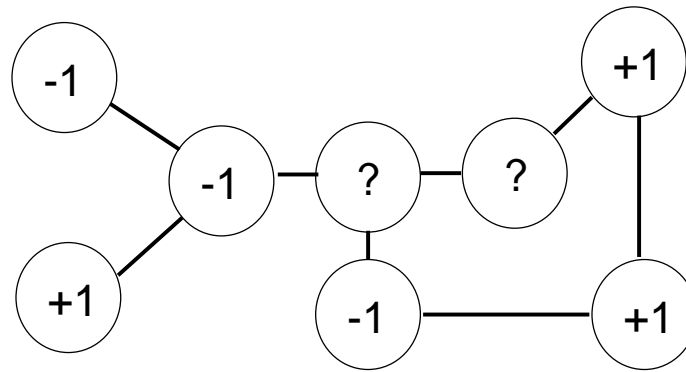
---

- Protein Interaction Network (Von Mering et al., Nature, 2002)
  - Edges represent physical interaction of proteins
  - Proteins with common functions often interact with each other
  - Identified via biological experiments
    - \* High-throughput yeast two hybrid, Correlated mRNA expression, Genetic interaction (synthetic letharity), Tandem affinity purification, High-throughput mass-spectrometric protein complex identification
  - 2617 proteins and 11855 edges
  - 76 two-class problems (functions)

# Protein Function Prediction on Interaction Network

---

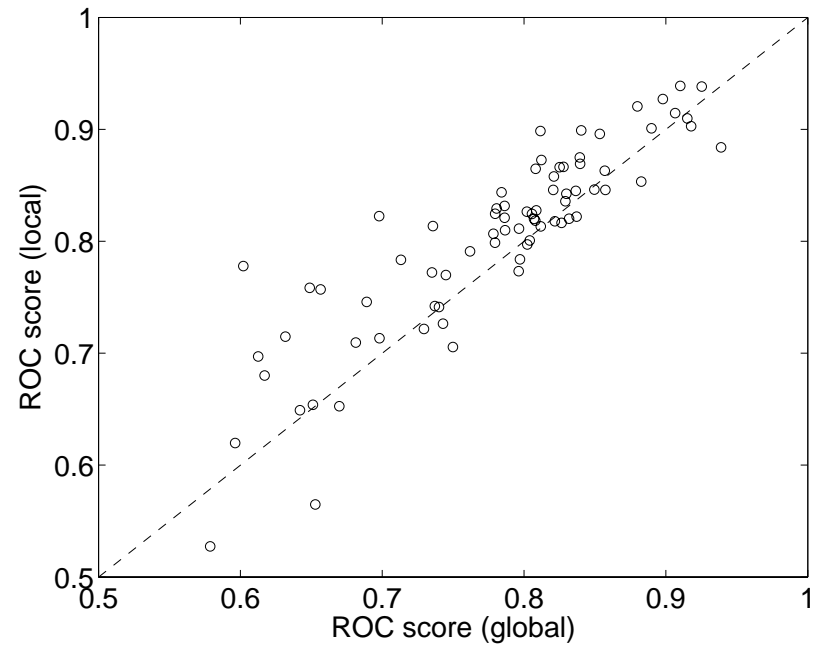
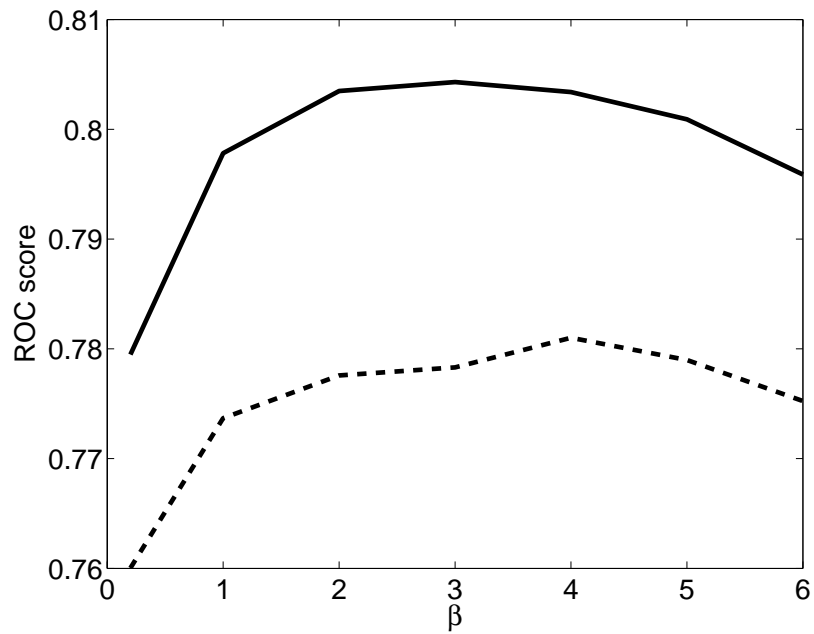
- Function prediction
  - +1,-1: Labeled proteins with/without a specific function
  - ?: Unlabeled proteins



# ROC Scores for Interaction Network

---

- SVM: 50% Training Nodes, 50% Testing Nodes



# Conclusion

---

- Diffusion kernel = Maximum von Neumann Entropy = e-projection
- Information geometry is a common framework
  - EM algorithm using von Neumann divergence
  - Matrix versions of other divergences inducing robustness (e.g.  $\beta$ -divergence)
  - Other ideas about constraints on kernel matrix
    - \* Partially known class labels