KRONECKER KERNEL RIDGE REGRESSION

- In MTP with target features, we often use kernel methods.
- Consider the following pairwise model representation in the primal:

$$f(\mathbf{x}, \mathbf{t}) = \boldsymbol{\omega}^\top (\phi(\mathbf{x}) \otimes \psi(\mathbf{t})),$$

  where ϕ is feature mapping for features and ψ is feature mapping for target (features) and $\otimes$ is Kronecker product.
- This yields Kronecker product pairwise kernel in the dual:

$$f(\mathbf{x}, \mathbf{t}) = \sum_{(\mathbf{x}', \mathbf{t}') \in \mathcal{D}} \alpha_{(\mathbf{x}', \mathbf{t}')} \cdot k(\mathbf{x}, \mathbf{x}') \cdot g(\mathbf{t}, \mathbf{t}') = \sum_{(\mathbf{x}', \mathbf{t}') \in \mathcal{D}} \alpha_{(\mathbf{x}', \mathbf{t}')} \Gamma((\mathbf{x}, \mathbf{t}), (\mathbf{x}', \mathbf{t}')),$$

  where k is kernel for feature map ϕ, g kernel for feature map ψ and $\alpha_{(\mathbf{x}', \mathbf{t}')}$ are dual parameters determined by:

$$\min_{\boldsymbol{\alpha}} ||\boldsymbol{\Gamma}\boldsymbol{\alpha} - \boldsymbol{z}||_2^2 + \lambda \boldsymbol{\alpha}^\top \boldsymbol{\Gamma}\boldsymbol{\alpha}, \text{ where } \boldsymbol{z} = \mathrm{vec}(Y)$$

- Commonly used in zero-shot learning.

Stock et al., A comparative study of pairwise learning methods based on kernel ridge regression, Neural Computation 2018.

PROBABILISTIC CLASSIFIER CHAINS

- Estimate the joint conditional distribution $\mathbb{P}(\mathbf{y} \mid \mathbf{x})$.
- For optimizing the subset 0/1 loss:

$$L_{0/1}(\mathbf{y}, \hat{\mathbf{y}}) = \mathbb{1}_{[\mathbf{y} \neq \hat{\mathbf{y}}]}$$

- Repeatedly apply the *product rule* of probability:

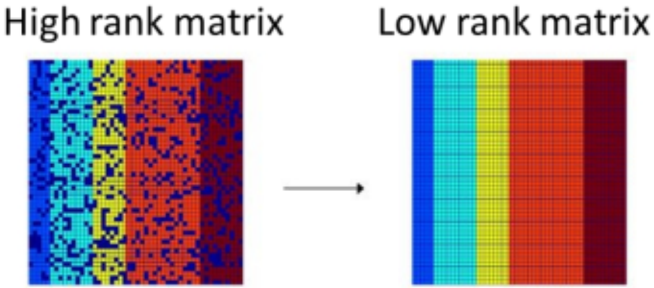
$$\mathbb{P}(\mathbf{y} \mid \mathbf{x}) = \prod_{j=m}^{l} \mathbb{P}(y_m \mid \mathbf{x}, y_1, \ldots, y_{m-1}) \,.$$

- Learning relies on constructing probabilistic classifiers for

$$\mathbb{P}(y_m | \mathbf{x}, y_1, \ldots, y_{m-1}) \,,$$

independently for each $m = 1, \ldots, l$.

LOW-RANK APPROXIMATION



- Low rank = some structure is shared across targets
- Typically perform low-rank approx of param matrix:

$$\min_{\Theta} \|Y - \Phi\Theta\|_F^2 + \lambda\,\mathrm{rank}(\Theta)$$

Chen et al., A convex formulation for learning shared structures from multiple tasks, ICML 2009.