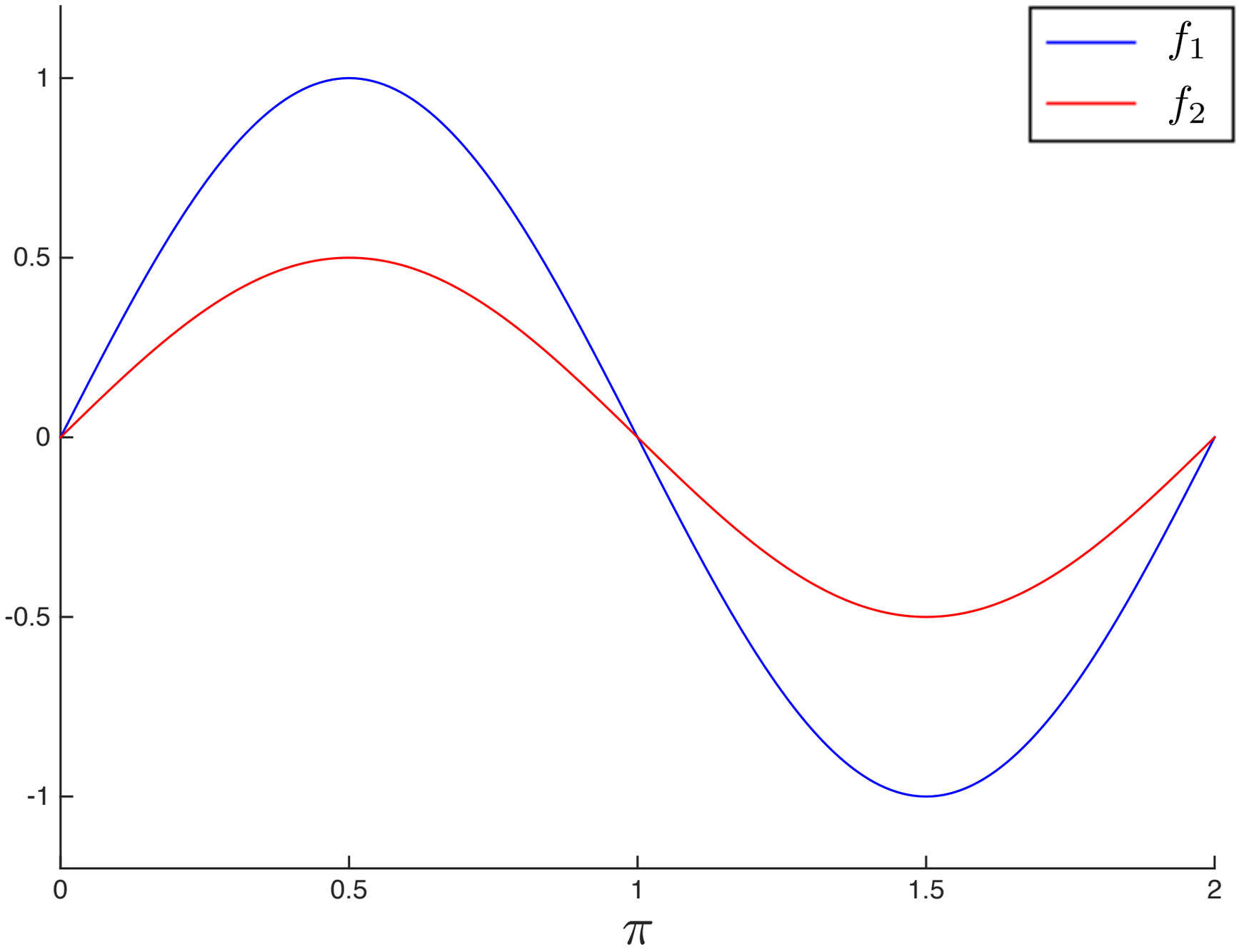
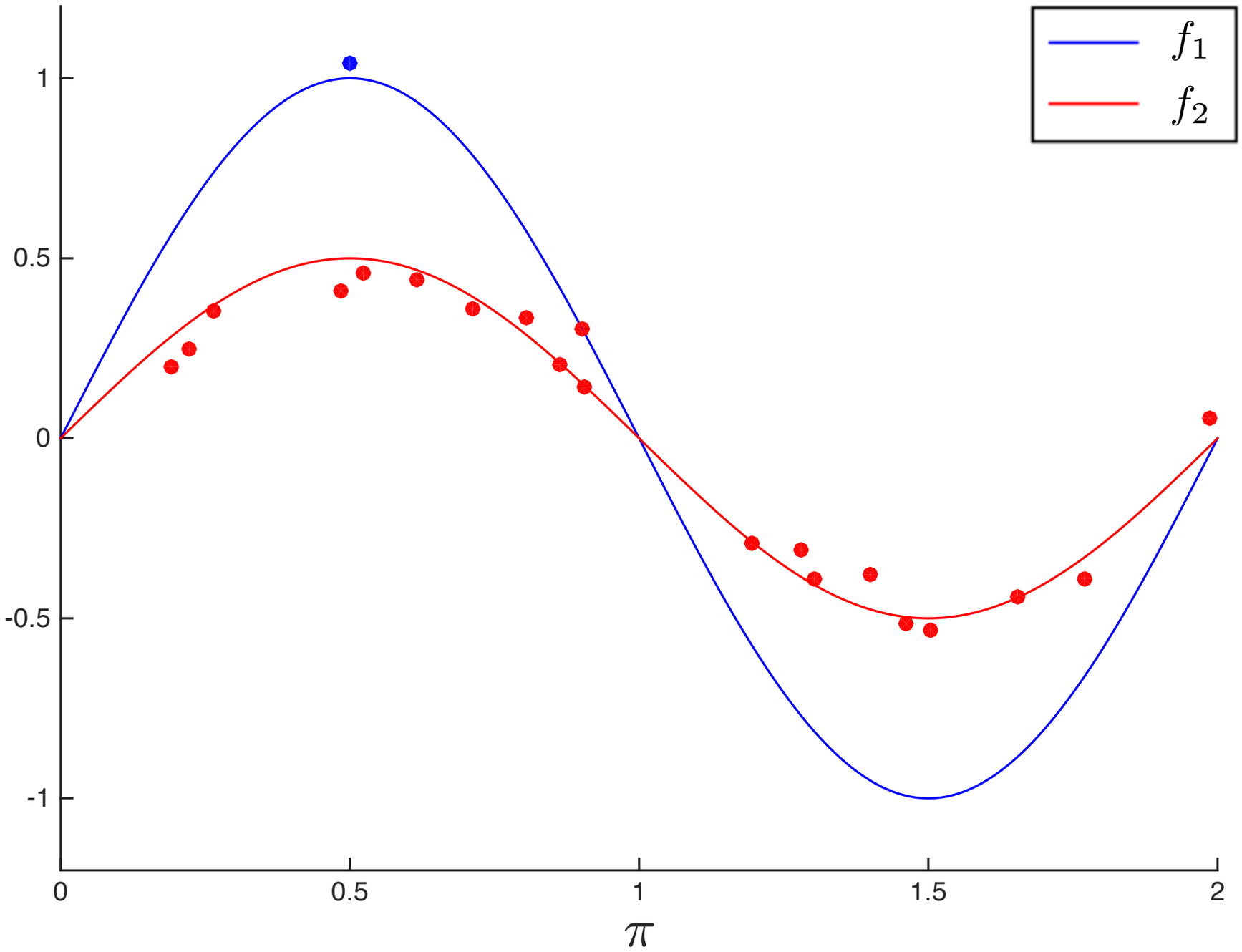
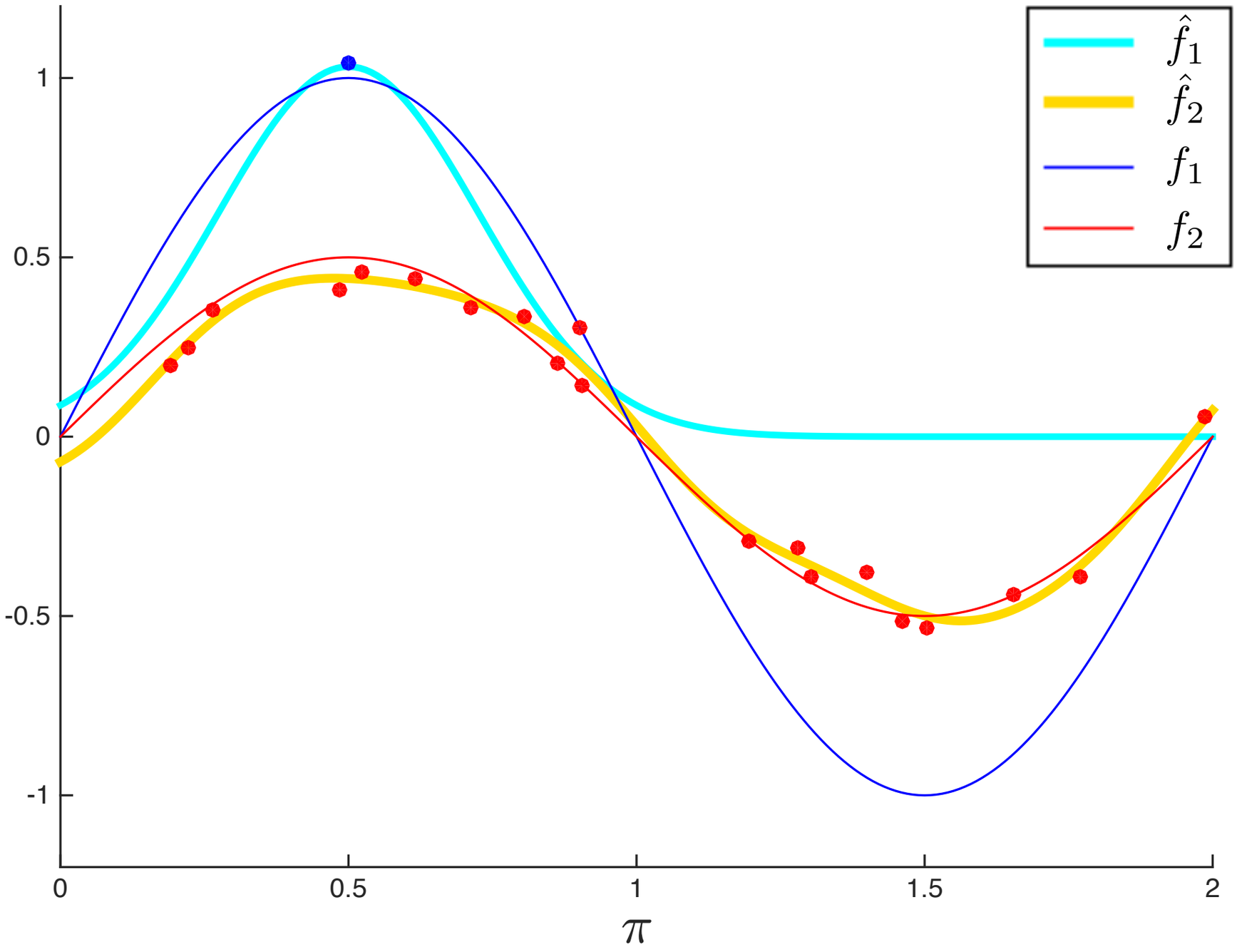
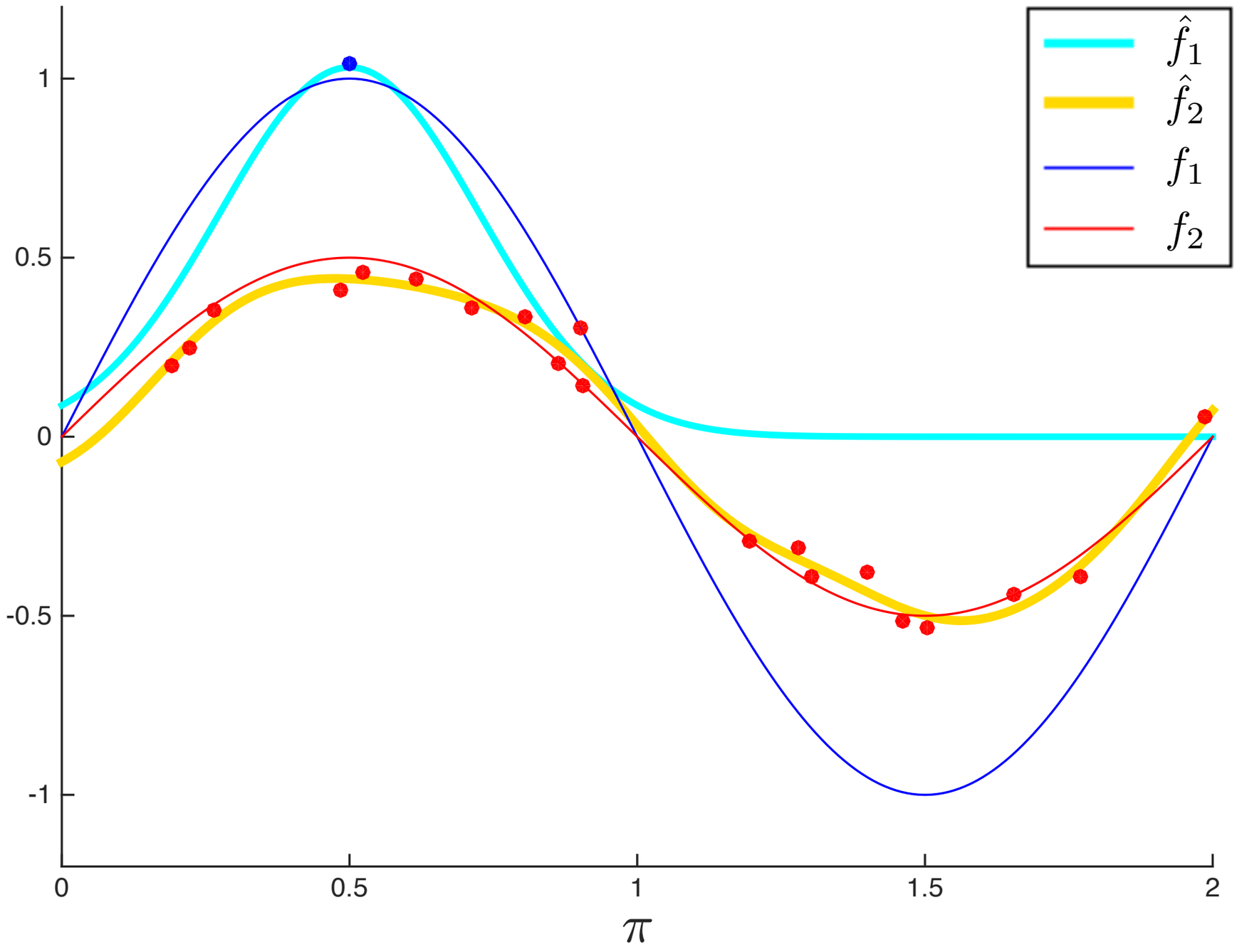
Convex Learning of Multiple Tasks and their Structure

Carlo Ciliberto,
Youssef Mroueh, Tomaso Poggio and Lorenzo Rosasco
Laboratory for Computational and Statistical Learning - Istituto Italiano di Tecnologia.
Dipartimento di Informatica Bioingegneria Robotica e Ingegneria dei Sistemi - Universita' di Genova
Poggio Lab - Massachusetts Institute of Technology.

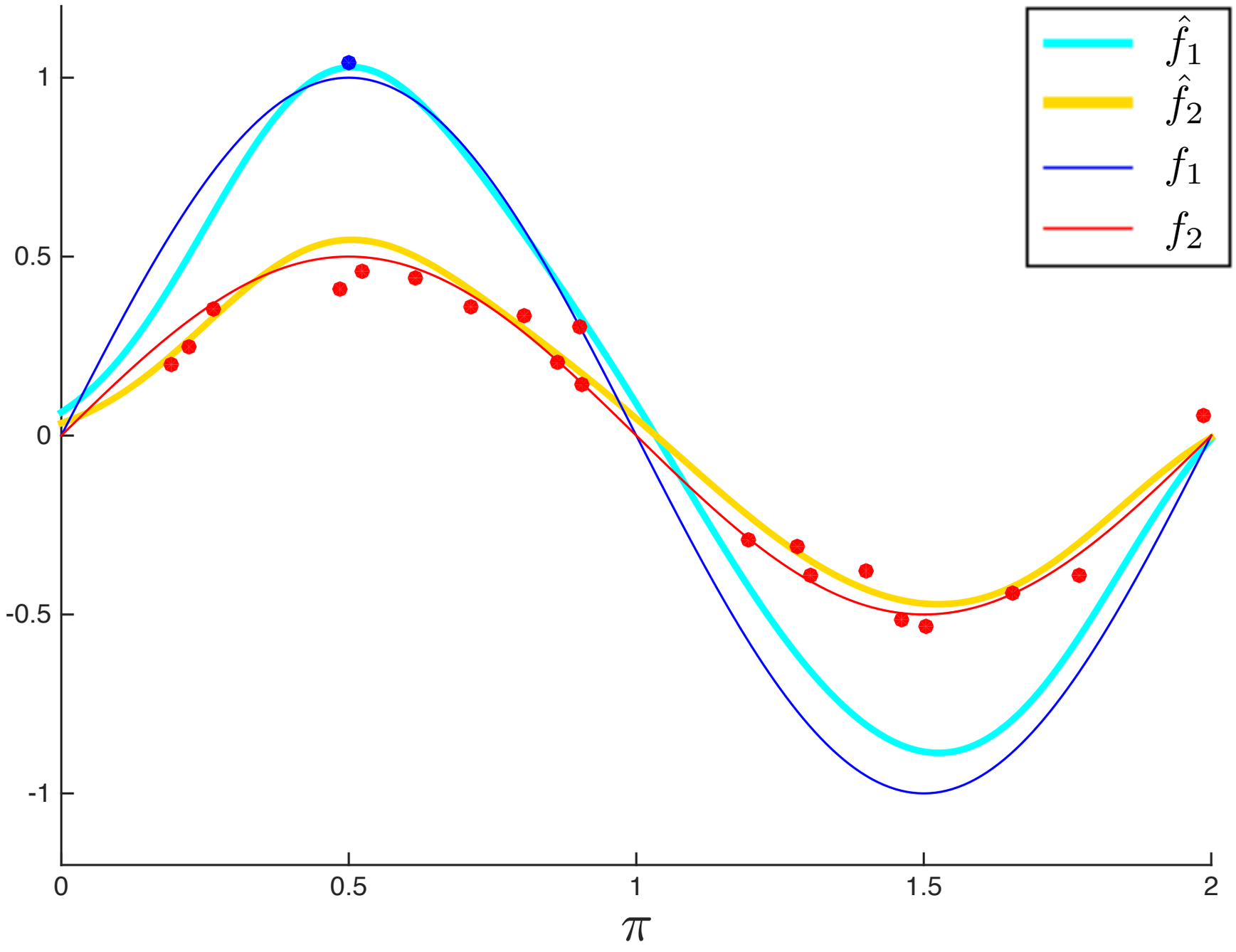

f_1
f_2
1
0.5
0
-0.5
-1
0
0.5
1
1.5
2
π


f_1
f_2
1
0.5
0
-0.5
-1
0
0.5
1
1.5
2
π


1
0.5
0
-0.5
-1
0
0.5
1
1.5
2
π
$\hat{f}_1$
$\hat{f}_2$
f_1
f_2



Assumption: f_1 and f_2 have "***similar***" behavior



Assumption: f_1 and f_2 have "***similar***" behavior

“Can we leverage on task-structure to reduce the supervision requirements?”

“Can we learn this structure when we don’t know it?”

Conjoint Analysis

Transfer Learning

Learning To Learn

Embeddings

Cluster-Task Learning

Task-relations Learning

Multi-Task Learning

Regularization

Structured Output Learning

Output Covariance Learning

Metric Deformation

Output Representation Learning

Collaborative Filtering

Co-Kriging

Outline

- Motivations
- Reproducing kernels Hilbert spaces for vector-valued functions
- A unifying framework for learning multiple tasks and their structure
- Recovering previous examples from the literature
- A novel sparse relations learning framework
- Summary

Outline

- Motivations
- **Reproducing kernels Hilbert spaces for vector-valued functions**
- A unifying framework for learning multiple tasks and their structure
- Recovering previous examples from the literature
- A novel sparse relations learning framework
- Summary

Tikhonov Regularization (scalar setting)

Training set $\{(x_i, y_i)\}_{i=1}^n$ with $x_i \in \mathcal{X}$ input space $y_i \in \mathbb{R}$ output space

Regularized problem

$$\underset{f \in \mathcal{H}}{\text{minimize}} \ \frac{1}{n} \sum_{i=1}^{n} \mathcal{L}(y_i, f(x_i)) + \lambda \| f \|_{\mathcal{H}}^2$$

reproducing kernel Hilbert space (RKHS) of scalar functions

Tikhonov Regularization (multi-task) [Micchelli and Pontil '04]

Training sets $\{\{(x_{it}, y_{it})\}_{i=1}^{n_t}\}_{t=1}^{T}$

Regularized problems

$$\underset{f_1,\ldots,f_T \in \mathcal{H}}{\text{minimize}} \; \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \sum_{t=1}^{T} \|f_t\|_{\mathcal{H}}^2$$

Tikhonov Regularization (multi-task) [Micchelli and Pontil '04]

Training sets $\{\{(x_{it}, y_{it})\}_{i=1}^{n_t}\}_{t=1}^{T}$

Idea: model the tasks as components of a ***vector-valued*** *function*

$$f = (f_1, \ldots, f_T) \in \mathcal{H}$$

Reproducing Kernel Hilbert Space of Vector-valued Functions

Tikhonov Regularization (multi-task) [Micchelli and Pontil '04]

Training sets $\{\{(x_{it}, y_{it})\}_{i=1}^{n_t}\}_{t=1}^{T}$

Regularized problems

$$\underset{f_1,\ldots,f_T \in \mathcal{H}}{\text{minimize}} \; \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \sum_{t=1}^{T} \|f_t\|_{\mathcal{H}}^2$$

Tikhonov Regularization (multi-task) [Micchelli and Pontil '04]

Training sets $\{\{(x_{it}, y_{it})\}_{i=1}^{n_t}\}_{t=1}^{T}$

Regularized problems

$$\underset{f \in \mathcal{H}}{\text{minimize}} \; \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \|f\|_{\mathcal{H}}^2$$

reproducing kernel Hilbert space of vector-valued functions (RKHSvv)

RKHS for vector-valued functions

$\mathcal{H}$ Hilbert space with inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$

$\Gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{T \times T}$ Positive definite ***Kernel*** s.t. $\forall x \in \mathcal{X} \quad \forall f \in \mathcal{H} \quad \forall c \in \mathcal{Y}$

Reproducing Property $\Gamma(x, \cdot)c \in \mathcal{H}$ and $\langle f, \Gamma(x, \cdot)c \rangle_{\mathcal{H}} = \langle f(x), c \rangle_{\mathcal{Y}}$

Representer Theorem

The ***regularized*** (multi-task) learning problem

$$\underset{f \in \mathcal{H}}{minimize} \;\; \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n_t} \sum_{i=1}^{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2$$

has solution $f_*(\cdot) = \sum_{\substack{t=1,n_t \\ i=1}} \Gamma(x_{it}, \cdot) c_{it} \in \mathcal{H}$ for suitable $c_{it} \in \mathbb{R}^T$

Separable Kernels

$$\Gamma(x, x') = k(x, x')A$$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

Scalar
Kernel

$A \in S_+^T$

Positive Semidefinite
(Encoding tasks **relations**)

Separable Kernels

$$\Gamma(x, x') = k(x, x')A$$

$k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$

Scalar
Kernel

$A \in S_+^T$

Positive Semidefinite
(Encoding tasks **relations**)

Advantages

I. Easy to construct from the scalar "experience"
II. Optimization (e.g. Sylvester method for least-squares [Mihn '11])
III. **Structure: k and A are enforcing structure on respectively the input and output spaces.**

Separable Kernels

$$\Gamma(x, x') = k(x, x')A$$

RKHS associated to k

Representer Theorem

$$f(\cdot) = \sum_{i=1}^{n} k(x_i, \cdot)Ac_i \in \mathcal{H} \qquad f_t(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i)A_t^\top c_i \in \mathcal{H}_k$$

input and output spaces.

Separable Kernels

$$\Gamma(x, x') = k(x, x')A$$

Representer Theorem

$$f(\cdot) = \sum_{i=1}^{n} k(x_i, \cdot)Ac_i \in \mathcal{H} \qquad f_t(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i)A_t^\top c_i \in \mathcal{H}_k$$

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j}^{n} k(x_i, x_j)c_i^\top Ac_j$$

Input and output spaces.

Separable Kernels

$$\Gamma(x, x') = k(x, x')A$$

Representer Theorem

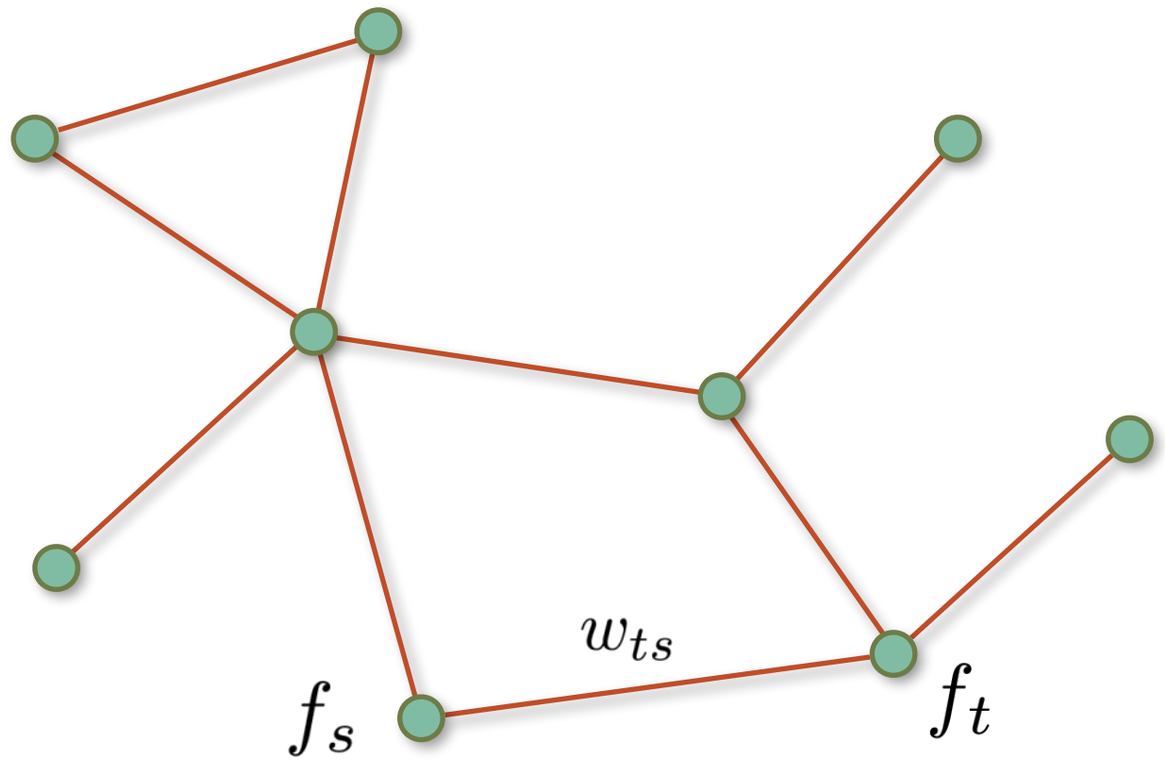
$$f(\cdot) = \sum_{i=1}^{n} k(x_i, \cdot)Ac_i \in \mathcal{H} \qquad f_t(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i)A_t^\top c_i \in \mathcal{H}_k$$

$$\|f\|_{\mathcal{H}}^2 = \sum_{i,j}^{n} k(x_i, x_j)c_i^\top Ac_j \longrightarrow \|f\|_{\mathcal{H}}^2 = \sum_{t,s}^{T} A_{ts}^\dagger \langle f_t, f_s \rangle_k$$

[Micchelli and Pontil '04]

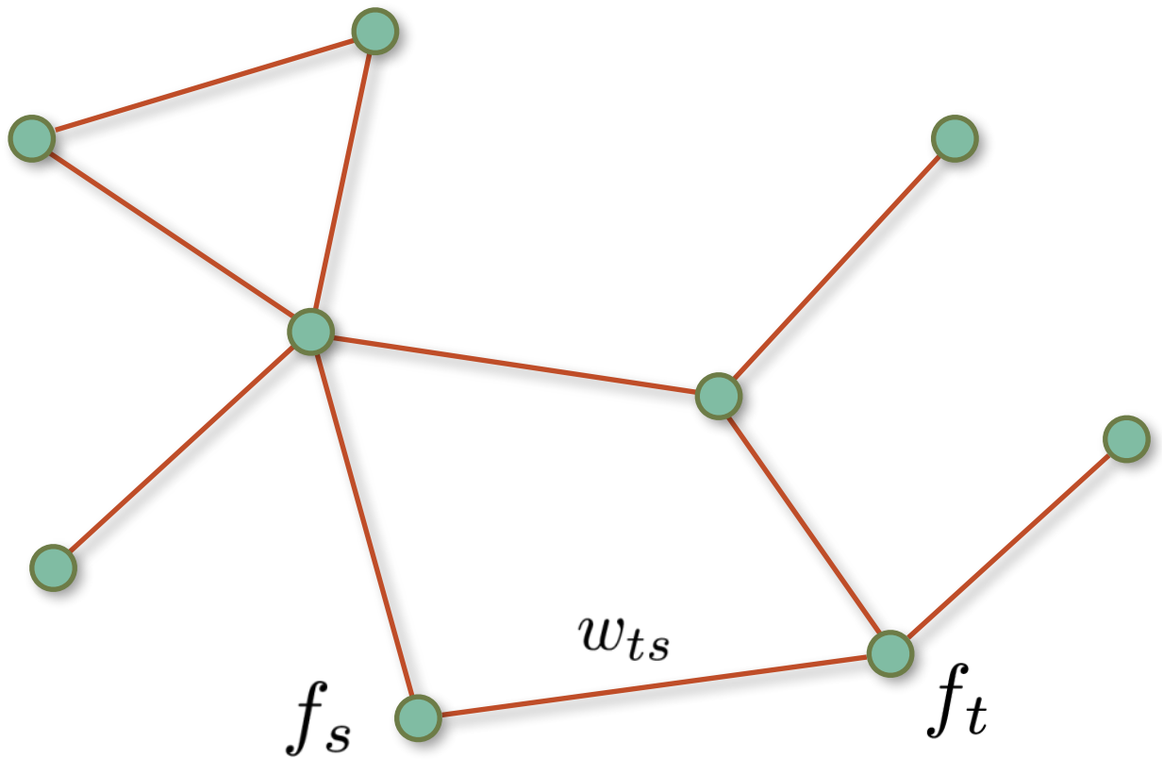
input and output spaces.

Known Tasks Similarity



[Micchelli and Pontil ’04]

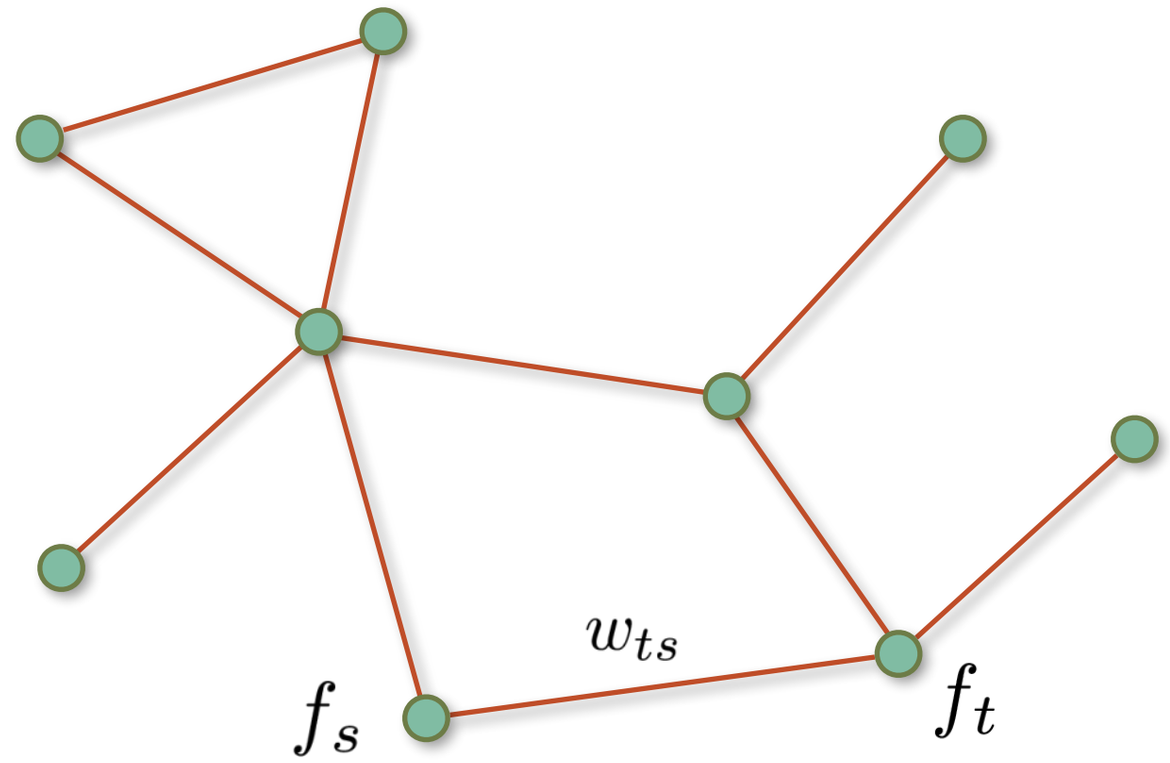
Known Tasks Similarity



$$\underset{f_1,\ldots,f_T \in \mathcal{H}_k}{minimize} \frac{1}{T}\sum_{\substack{t=1\\i=1}}^{T,n_t}\frac{1}{n_t}\mathcal{L}(y_{it}, f_t(x_{it}))+\lambda\sum_{t,s}^{T} w_{ts}\|f_t - f_s\|_k^2 + \gamma\sum_{t}^{T}\|f_t\|_k^2$$

[Micchelli and Pontil '04]

Known Tasks Similarity



$$\underset{f \in \mathcal{H}}{minimize} \; \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2$$

$$A^{\dagger} = L_W + \frac{\gamma}{\lambda} I$$

L_W Graph Laplacian

[Micchelli and Pontil '04]

“Can we leverage on task-structure to reduce the supervision requirements?”

“Can we learn this structure when we don’t know it?”

Outline

- Motivations
- Reproducing kernels Hilbert spaces for vector-valued functions
- **A unifying framework for learning multiple tasks and their structure**
- Recovering previous examples from the literature
- A novel sparse relations learning framework
- Summary

$$\underset{f \in \mathcal{H}}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T,n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2$$

$$\underset{f \in \mathcal{H} \quad A \in S_+^T}{minimize} \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T,n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2$$

$$\underset{f \in \mathcal{H} \quad A \in S_+^T}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

"structure" penalty

$$\underset{\substack{f \in \mathcal{H} \quad A \in S_+^T}}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T,n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \|f\|_{\mathcal{H}}^2 + F(A)$$

"structure" penalty

Examples

$F(A) = i_{\{\mathrm{tr}(A) \leq 1\}}$ trace constraint [Argyriou '08] (more on this later…)

$F(A) = i_{\{\mathrm{tr}(A) = 1\}}$ trace constraint [Zhang '10]

$F(A) = \|A\|_F^2$ squared Frobenius norm [Dinuzzo '11]

$F(A) = \|A\|_p^k$ any k-power of a p-Schatten norm

$F(A) = \|A\|_{\mathrm{CN}}$ "cluster" norm [Jacob et al. '09]

$F(A) = \|A\|_{\ell_1}$ sum of absolute values [Ciliberto '15]

$$\underset{f \in \mathcal{H} \quad A \in S_+^T}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T,n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

$$f(x) = \sum_{i=1}^{n} k(x, x_i) A c_i$$

$$\| f \|_{\mathcal{H}}^2 = \sum_{i,j}^{n} k(x_i, x_j) c_i^\top A c_j$$

$K \in S_+^n$ Kernel Matrix

$Y \in \mathbb{R}^{n \times T}$ Output Matrix

$$\underset{f \in \mathcal{H} \quad A \in S_+^T}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T,n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

$$f(x) = \sum_{i=1}^{n} k(x, x_i) A c_i$$

$$\| f \|_{\mathcal{H}}^2 = \sum_{i,j}^{n} k(x_i, x_j) c_i^\top A c_j$$

$K \in S_+^n$ Kernel Matrix

$Y \in \mathbb{R}^{n \times T}$ Output Matrix

$$\underset{C \in \mathbb{R}^{n \times T} \; A \in S_+^T}{minimize} \quad \mathcal{L}(Y, KCA) + \lambda \; tr(AC^\top KC) + F(A) \qquad (\mathcal{Q})$$

$$\underset{f \in \mathcal{H} \quad A \in S_+^T}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T,n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

$$f(x) = \sum_{i=1}^{n} k(x, x_i) A c_i$$

$$\| f \|_{\mathcal{H}}^2 = \sum_{i,j}^{n} k(x_i, x_j) c_i^\top A c_j$$

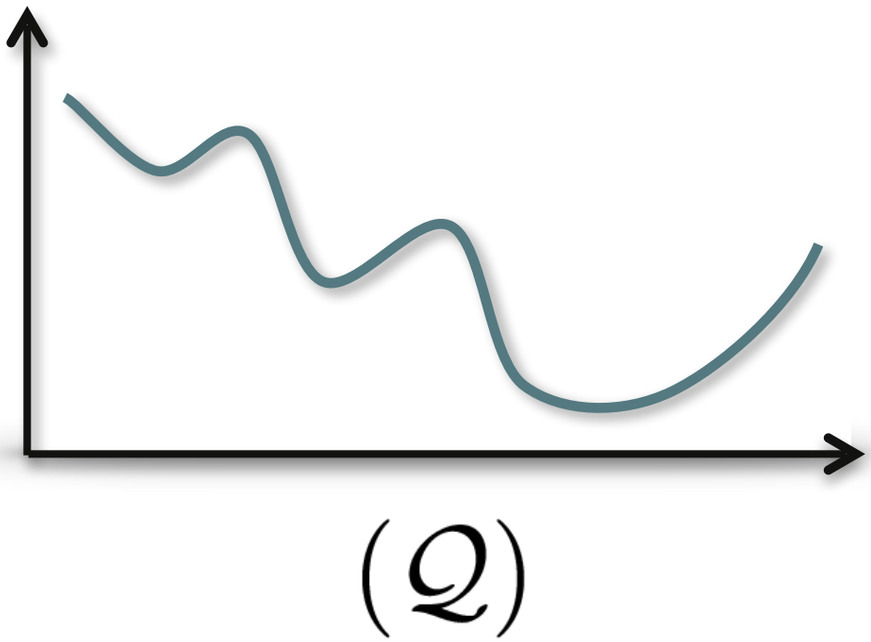
$K \in S_+^n$ Kernel Matrix

$Y \in \mathbb{R}^{n \times T}$ Output Matrix

$$\underset{C \in \mathbb{R}^{n \times T} \; A \in S_+^T}{minimize} \quad \mathcal{L}(Y, KCA) + \lambda \; tr(AC^\top KC) + F(A) \qquad (\mathcal{Q})$$

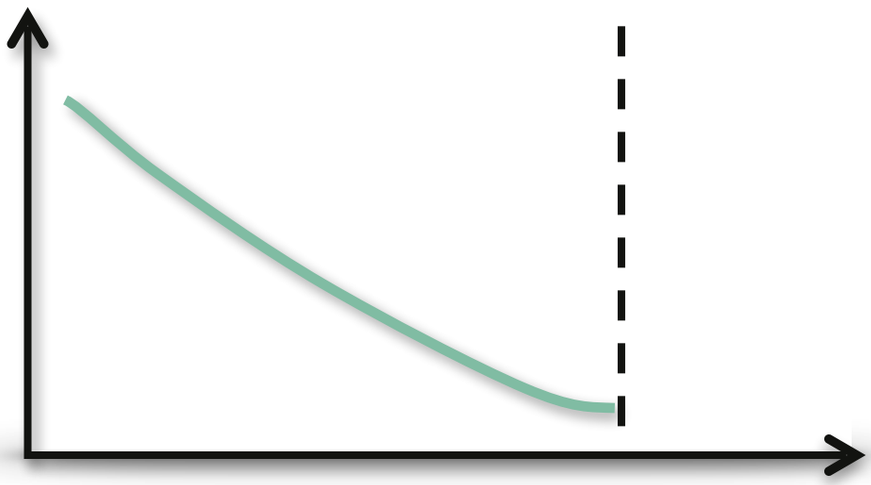
Non Convex!

Roadmap

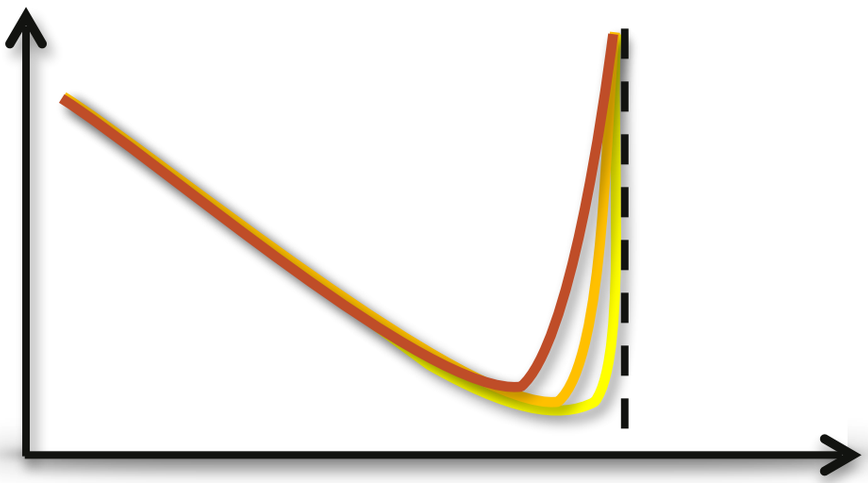


$(\mathcal{Q})$

Equivalence



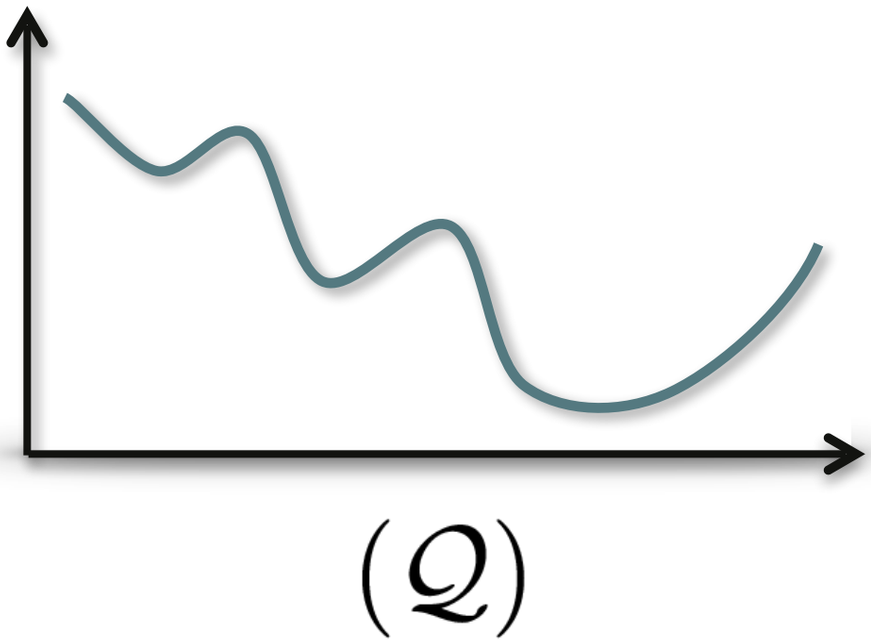
$(\mathcal{R})$



$(\mathcal{S}^{\delta})$

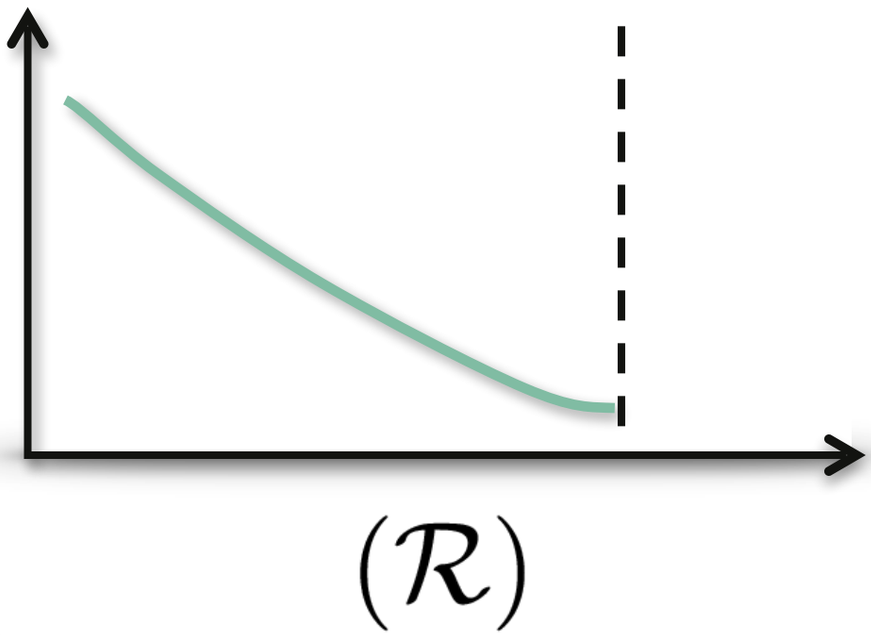
Barrier

Roadmap

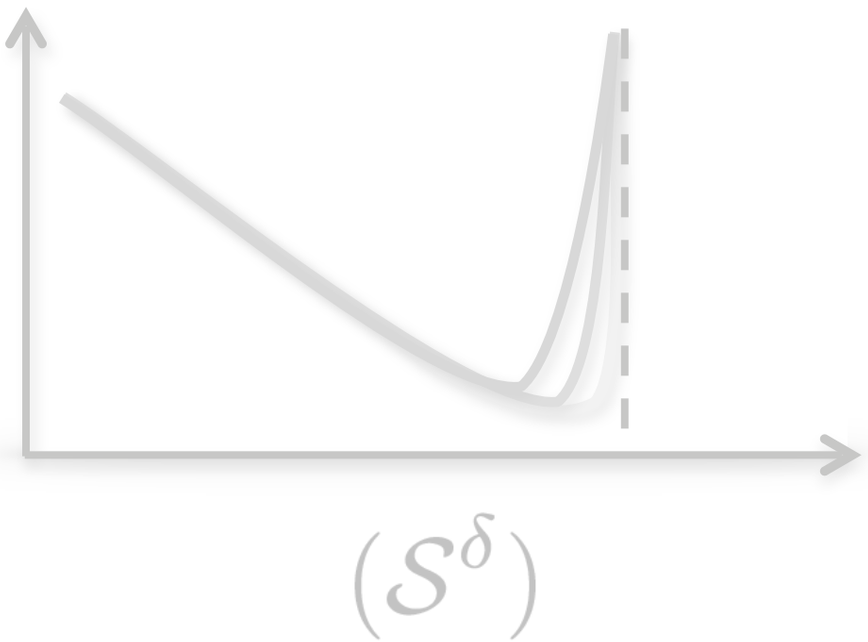


$(\mathcal{Q})$

Equivalence



$(\mathcal{R})$



$(\mathcal{S}^{\delta})$

Barrier

$$\underset{(B,A)\,\in\,\mathcal{B}}{minimize} \quad \mathcal{L}(Y, KB) + \lambda\ tr(A^{\dagger} B^{\top} K B) + F(A) \qquad (\mathcal{R})$$

pseudoinverse

$$\underset{(B,A)\in\mathcal{B}}{minimize} \quad \mathcal{L}(Y, KB) + \lambda\ tr(A^{\dagger} B^{\top} KB) + F(A) \qquad (\mathcal{R})$$

A is missing!

[Ciliberto et al. ’13]

$$\underset{(B,A)\,\in\,\mathcal{B}}{minimize} \quad \mathcal{L}(Y, KB) + \lambda\ tr(A^{\dagger} B^{\top} K B) + F(A) \qquad (\mathcal{R})$$

$$\mathcal{B} = \{(B, A) \in \mathbb{R}^{n \times T} \times S_{+}^{T} \mid \mathcal{R}(B^{\top} K B) \subseteq \mathcal{R}(A)\}$$

[Ciliberto et al. ’13]

$$\underset{(B,A)\,\in\,\mathcal{B}}{minimize} \quad \mathcal{L}(Y, KB) + \lambda\ tr(A^{\dagger}B^{\top}KB) + F(A) \qquad (\mathcal{R})$$

$$\mathcal{B} = \{(B, A) \in \mathbb{R}^{n\times T} \times S_{+}^{T} \mid \mathcal{R}(B^{\top}KB) \subseteq \mathcal{R}(A)\}$$

Theorem

1) $(Q) \iff (\mathcal{R})$

Solutions

$$(C_Q, A_Q) \longmapsto (C_Q A_Q, A_Q)$$

$$(B_R, A_R) \longmapsto (B_R A_R^{\dagger}, A_R)$$

2) $\mathcal{B}$ convex and $tr(A^{\dagger}B^{\top}KB)$ convex on $\mathcal{B}$

$$\underset{(B,A)\,\in\,\mathcal{B}}{minimize} \quad \mathcal{L}(Y, KB) + \lambda\, tr(A^{\dagger} B^{\top} K B) + F(A) \qquad (\mathcal{R})$$

$$\mathcal{B} = \{(B, A) \in \mathbb{R}^{n \times T} \times S_{+}^{T} \mid \mathcal{R}(B^{\top} K B) \subseteq \mathcal{R}(A)\}$$

Theorem

1) $(Q) \iff (\mathcal{R})$

Solutions

$$(C_Q, A_Q) \longmapsto (C_Q A_Q, A_Q)$$

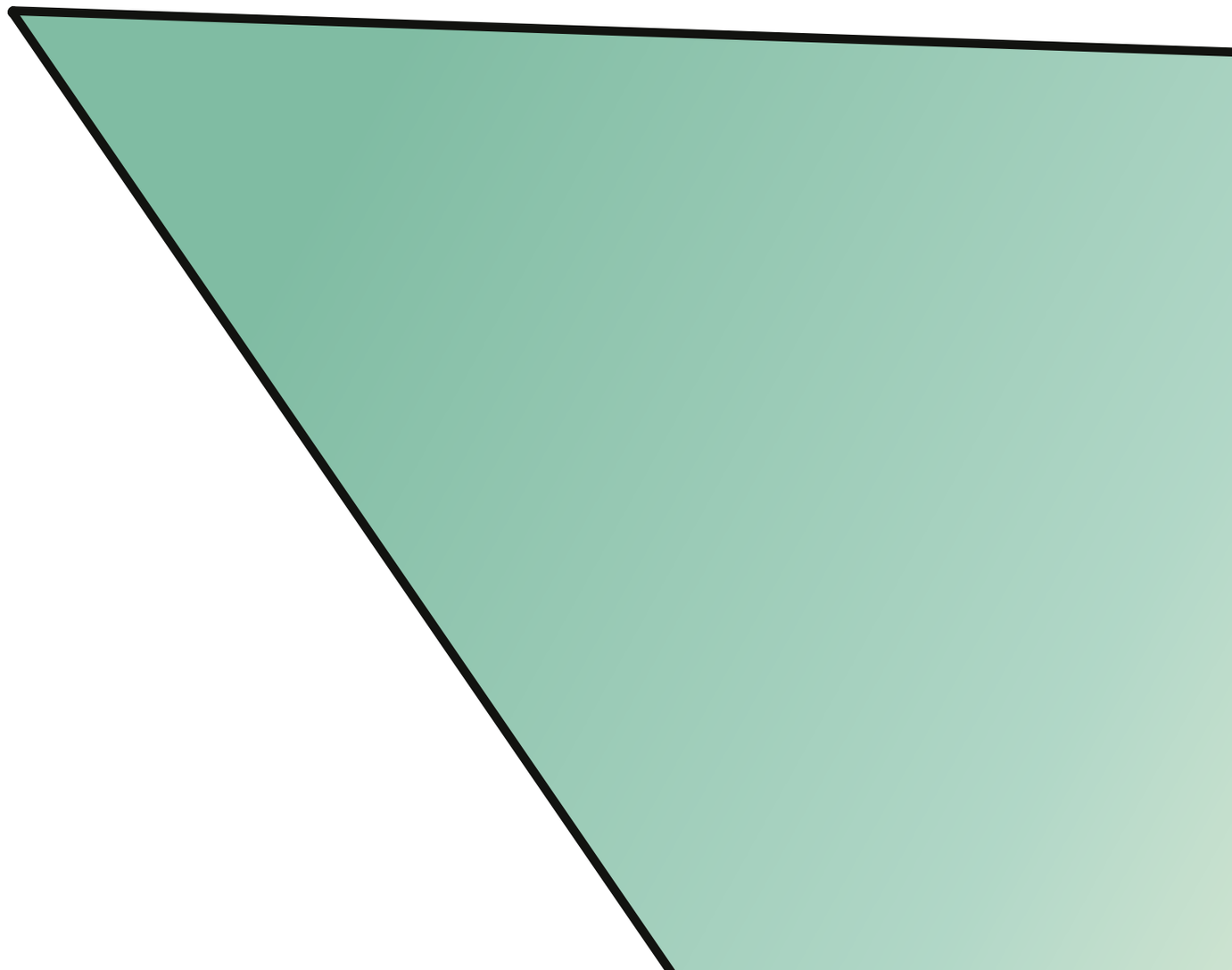
$$(B_R, A_R) \longmapsto (B_R A_R^{\dagger}, A_R)$$

2) $\mathcal{B}$ convex and $tr(A^{\dagger} B^{\top} K B)$ convex on $\mathcal{B}$

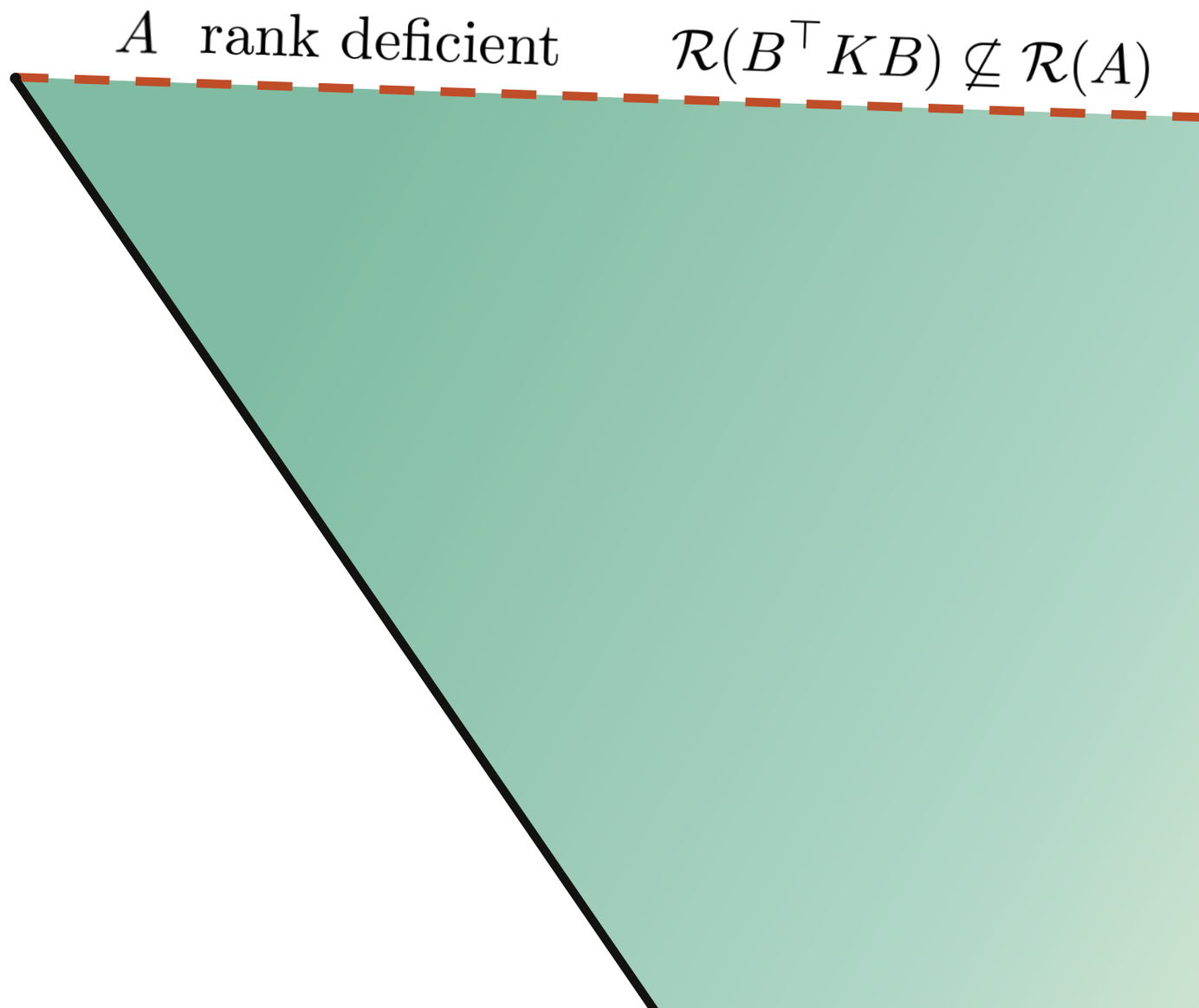
$$\mathcal{L}, F \text{ convex} \implies (\mathcal{R}) \text{ convex}$$

Generalization of Output Kernel Leraning [Dinuzzo et al. ’11]

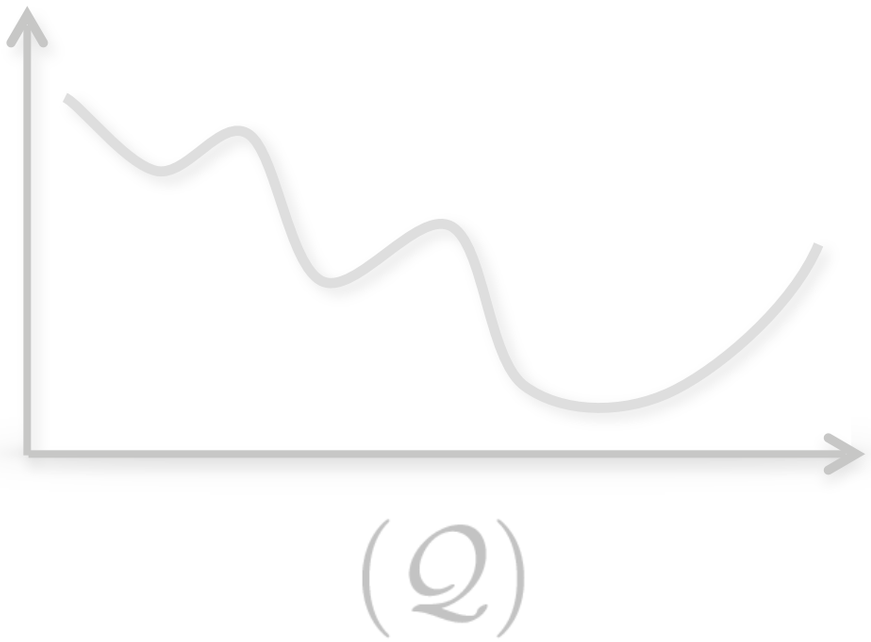
$$dom\mathcal{Q} \qquad \mathbb{R}^{n\times T} \times S_+^T$$



$$
\begin{array}{ll}
dom\mathcal{Q} & \mathbb{R}^{n\times T} \times S_+^T \\
 & \cup| \\
dom\mathcal{R} & \mathcal{B}
\end{array}
$$

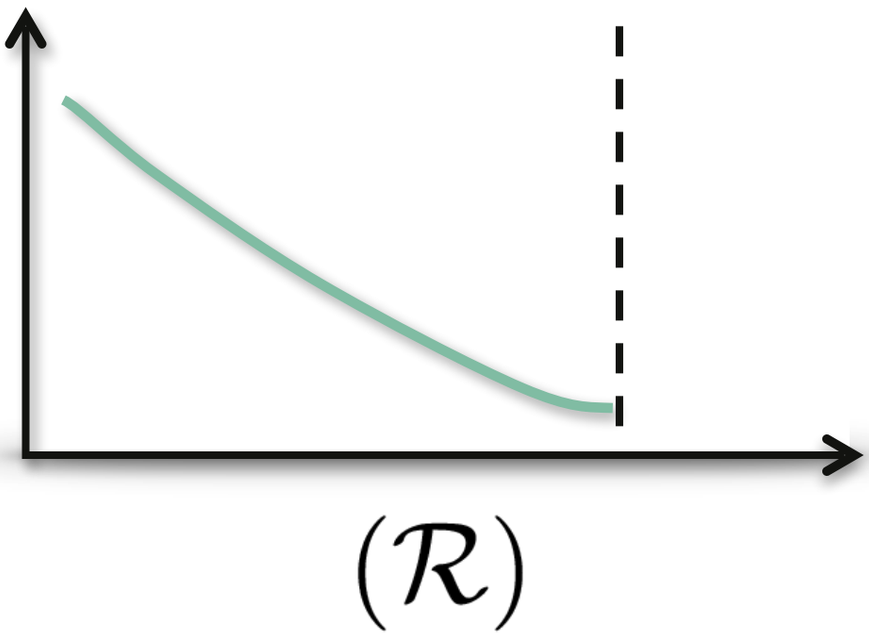


Roadmap

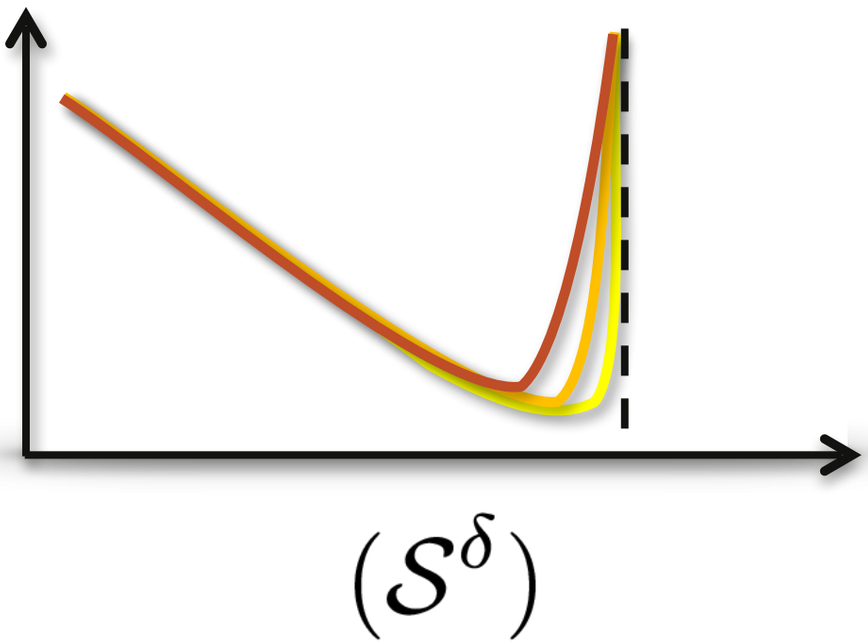


$(\mathcal{Q})$

Equivalence



$(\mathcal{R})$



$(\mathcal{S}^{\delta})$

Barrier

$$\underset{(B,A)\in\mathcal{B}}{minimize}\;\; \mathcal{L}(Y, KB) + \lambda\; tr(A^{\dagger}B^{\top}KB) + F(A) \qquad (\mathcal{R})$$

$$\underset{(B,A)\in\mathcal{B}}{minimize}\ \ \mathcal{L}(Y,KB)+\lambda\ tr(A^{\dagger}B^{\top}KB)+F(A)+\delta^{2}tr(A^{-1})$$

Barrier

idea from [Argyriou et al. 08]

$$\underset{B \in \mathbb{R}^{n \times T} A \in S_+^T}{\textit{minimize}} \; \mathcal{L}(Y, KB) + \lambda \; tr(A^{-1}(B^\top K B + \delta^2 I)) + F(A) \qquad (\mathcal{S}^\delta)$$

$$\underset{B \in \mathbb{R}^{n \times T}\, A \in S_+^T}{minimize}\ \mathcal{L}(Y, KB) + \lambda\ tr(A^{-1}(B^\top K B + \delta^2 I)) + F(A) \qquad (\mathcal{S}^\delta)$$

Theorem

Given $\delta_n \longrightarrow 0$ for $n \longrightarrow +\infty$

$$\mathcal{S}_*^{\delta_n} = \underset{B \in \mathbb{R}^{n \times T}\ A \in S_+^T}{argmin}\ S^{\delta_n}(B, A) \qquad \underset{n \longrightarrow +\infty}{\longrightarrow} \qquad R_* = \underset{(B,A) \in \mathcal{B}}{argmin}\ R\ (B, A)$$

$$\underset{B \in \mathbb{R}^{n \times T}\, A \in S_+^T}{minimize}\ \mathcal{L}(Y, KB) + \lambda\ tr(A^{-1}(B^\top K B + \delta^2 I)) + F(A) \qquad (\mathcal{S}^\delta)$$

Theorem

Given $\delta_n \longrightarrow 0$ for $n \longrightarrow +\infty$

$$\mathcal{S}_*^{\delta_n} = \underset{B \in \mathbb{R}^{n \times T}\ A \in S_+^T}{argmin}\ S^{\delta_n}(B, A) \qquad \underset{n \longrightarrow +\infty}{\longrightarrow} \qquad R_* = \underset{(B, A) \in \mathcal{B}}{argmin}\ R\ (B, A)$$

$(\mathcal{S}^\delta)$ Is a good proxy for $(\mathcal{R})$

$$\underset{B \in \mathbb{R}^{n \times T} A \in S_+^T}{minimize} \; \mathcal{L}(Y, KB) + \lambda \; tr(A^{-1}(B^\top KB + \delta^2 I)) + F(A)$$

Algorithm 1 ALTERNATING MINIMIZATION

Input: K, Y, ϵ tolerance, δ perturbation parameter, S objective functional of $(\mathcal{S}^\delta)$, $\mathcal{L}$ loss, F structure penalty.

Initialize: choose $(B_0, A_0), t = 0$

repeat

$B_{t+1} \leftarrow$ SUPERVISEDSTEP $(\mathcal{L}, K, Y, A_t)$

$A_{t+1} \leftarrow$ UNSUPERVISEDSTEP(F, K, δ, B_{t+1})

$t \leftarrow t + 1$

until $|S(B_{t+1}, A_{t+1}) - S(B_t, A_t)| < \epsilon$

$$\underset{B \in \mathbb{R}^{n \times T}}{\underset{A \in S_+^T}{minimize}} \; \mathcal{L}(Y, KB) + \lambda \; tr(A^{-1}(B^\top KB + \delta^2 I)) + F(A)$$

Algorithm 1 ALTERNATING MINIMIZATION

Input: K, Y, ϵ tolerance, δ perturbation parameter, S objective functional of $(\mathcal{S}^\delta)$, $\mathcal{L}$ loss, F structure penalty.

Initialize: choose $(B_0, A_0), t = 0$

repeat

$B_{t+1} \leftarrow$ SUPERVISEDSTEP $(\mathcal{L}, K, Y, A_t)$

$A_{t+1} \leftarrow$ UNSUPERVISEDSTEP(F, K, δ, B_{t+1})

$t \leftarrow t + 1$

until $|S(B_{t+1}, A_{t+1}) - S(B_t, A_t)| < \epsilon$

$$\underset{B \in \mathbb{R}^{n \times T} \, A \in S_+^T}{minimize} \; \mathcal{L}(Y, KB) + \lambda \; tr(A^{-1}(B^\top K B + \delta^2 I)) + F(A)$$

Algorithm 1 ALTERNATING MINIMIZATION

Input: K, Y, ϵ tolerance, δ perturbation parameter, S objective functional of $(\mathcal{S}^\delta)$, $\mathcal{L}$ loss, F structure penalty.

Initialize: choose $(B_0, A_0), t = 0$

repeat

$B_{t+1} \leftarrow$ SUPERVISEDSTEP $(\mathcal{L}, K, Y, A_t)$

$A_{t+1} \leftarrow$ UNSUPERVISEDSTEP(F, K, δ, B_{t+1})

$t \leftarrow t + 1$

until $|S(B_{t+1}, A_{t+1}) - S(B_t, A_t)| < \epsilon$

$$\underset{B \in \mathbb{R}^{n \times T}}{\underset{A \in S_+^T}{minimize}} \; \mathcal{L}(Y, KB) + \lambda \; tr(A^{-1}(B^\top KB + \delta^2 I)) + F(A)$$

Algorithm 1 ALTERNATING MINIMIZATION

Input: K, Y, ϵ tolerance, δ perturbation parameter, S objective functional of $(\mathcal{S}^\delta)$, $\mathcal{L}$ loss, F structure penalty.

Initialize: choose $(B_0, A_0), t = 0$

repeat

$B_{t+1} \leftarrow$ SUPERVISEDSTEP $(\mathcal{L}, K, Y, A_t)$

$A_{t+1} \leftarrow$ UNSUPERVISEDSTEP(F, K, δ, B_{t+1})

$t \leftarrow t + 1$

until $|S(B_{t+1}, A_{t+1}) - S(B_t, A_t)| < \epsilon$

Corollary

(from Theorem 2 in [Razaviyayin et al. '13] or Theorem 4.2 in [Tseng '01])

Every limiting point of a minimizing sequence provided by ***Algoritm 1*** *is a* ***global*** *minimizer for problem* $(\mathcal{S}^{\delta})$

Algorithm 1 ALTERNATING MINIMIZATION

Input: K, Y, ϵ tolerance, δ perturbation parameter, S objective functional of $(\mathcal{S}^{\delta})$, $\mathcal{L}$ loss, F structure penalty.

Initialize: choose $(B_0, A_0), t = 0$

repeat

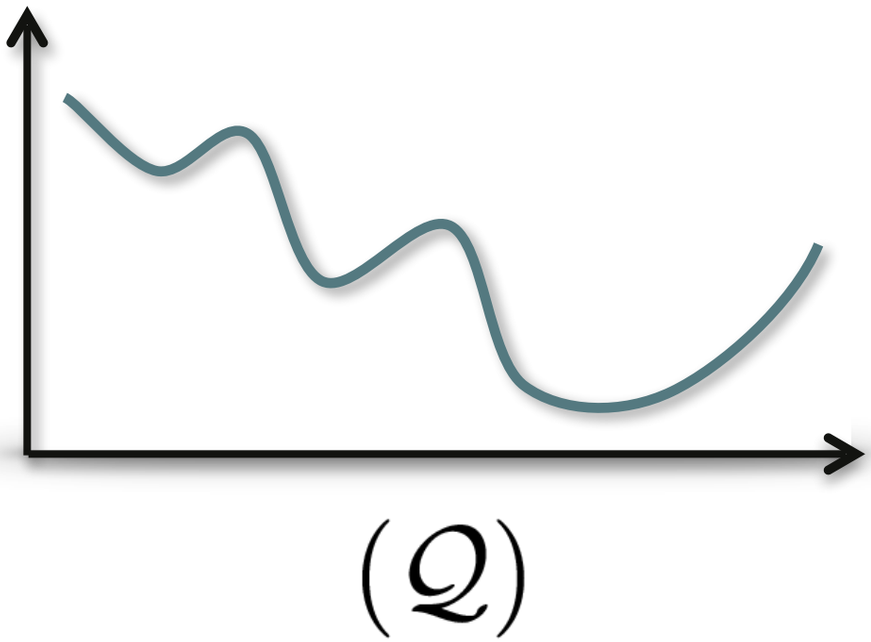
$B_{t+1} \leftarrow$ SUPERVISEDSTEP $(\mathcal{L}, K, Y, A_t)$

$A_{t+1} \leftarrow$ UNSUPERVISEDSTEP(F, K, δ, B_{t+1})

$t \leftarrow t + 1$

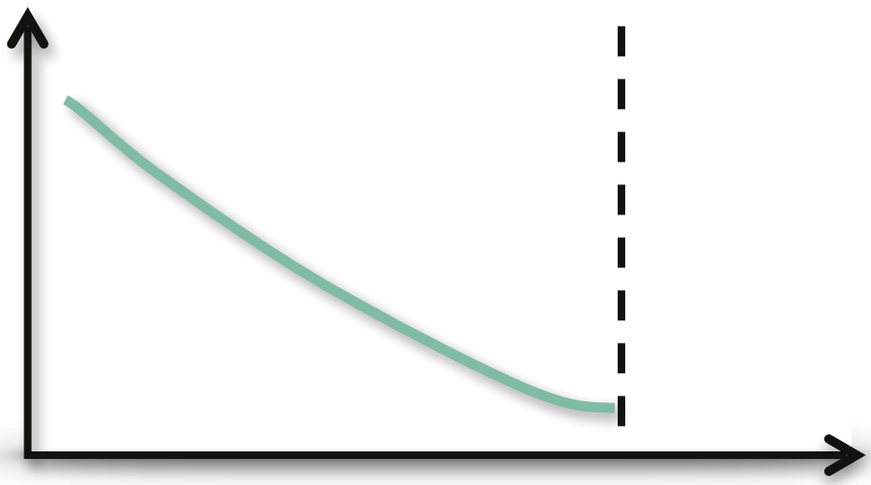
until $|S(B_{t+1}, A_{t+1}) - S(B_t, A_t)| < \epsilon$

Roadmap

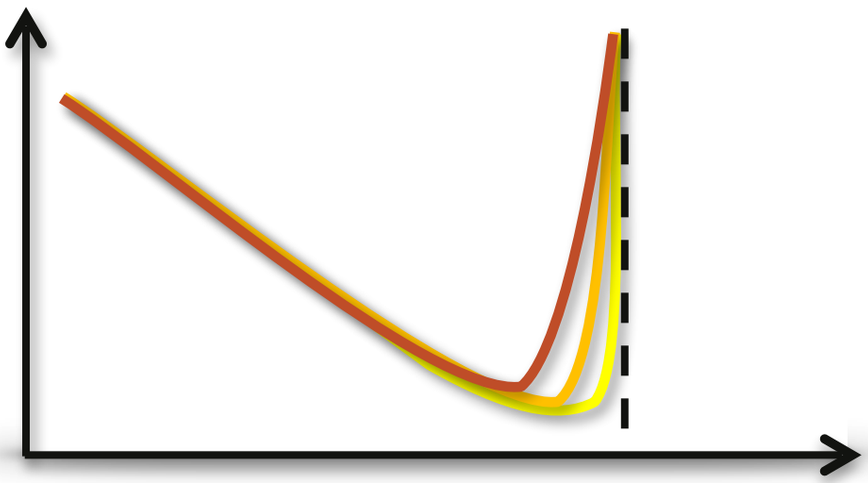


$(\mathcal{Q})$

Equivalence



$(\mathcal{R})$



$(\mathcal{S}^{\delta})$

Barrier

Recap

- General setting for learning multiple tasks and their ***structure***.

Recap

- General setting for learning multiple tasks and their ***structure***.

- *Equivalence with a **convex** learning problem.*

Recap

- General setting for learning multiple tasks and their ***structure***.

- *Equivalence with a **convex** learning problem.*

- ***Provably** solvable by block coordinate descent.*

Recap

- General setting for learning multiple tasks and their ***structure***.

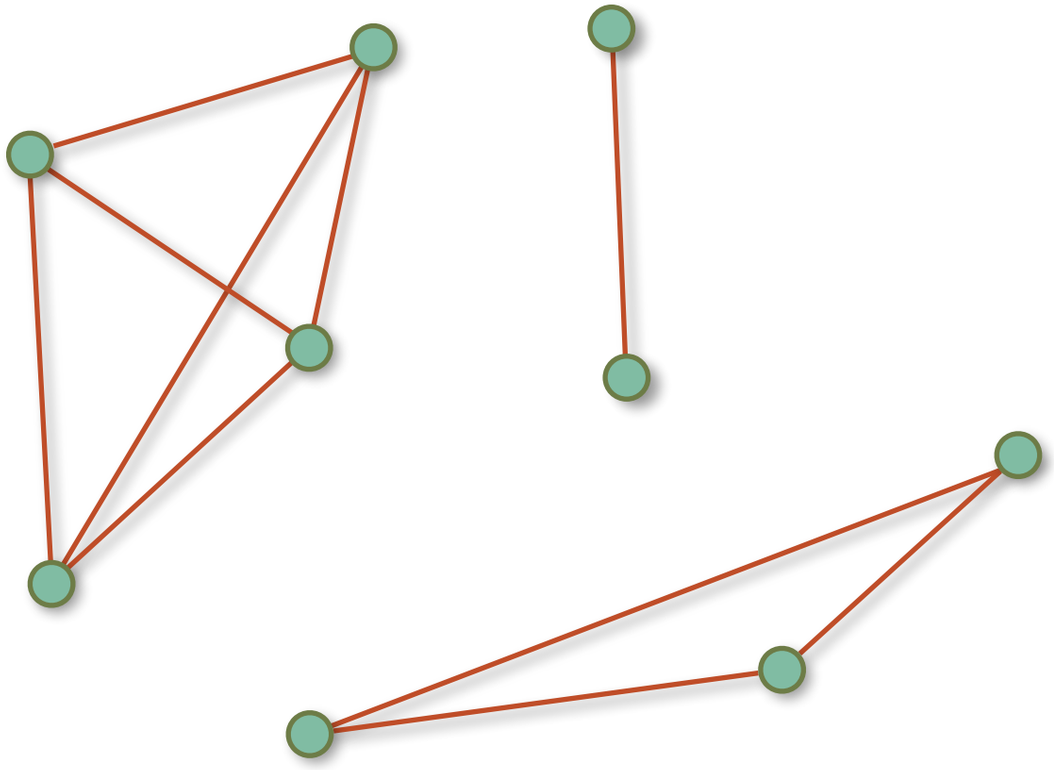
- *Equivalence with a **convex** learning problem.*

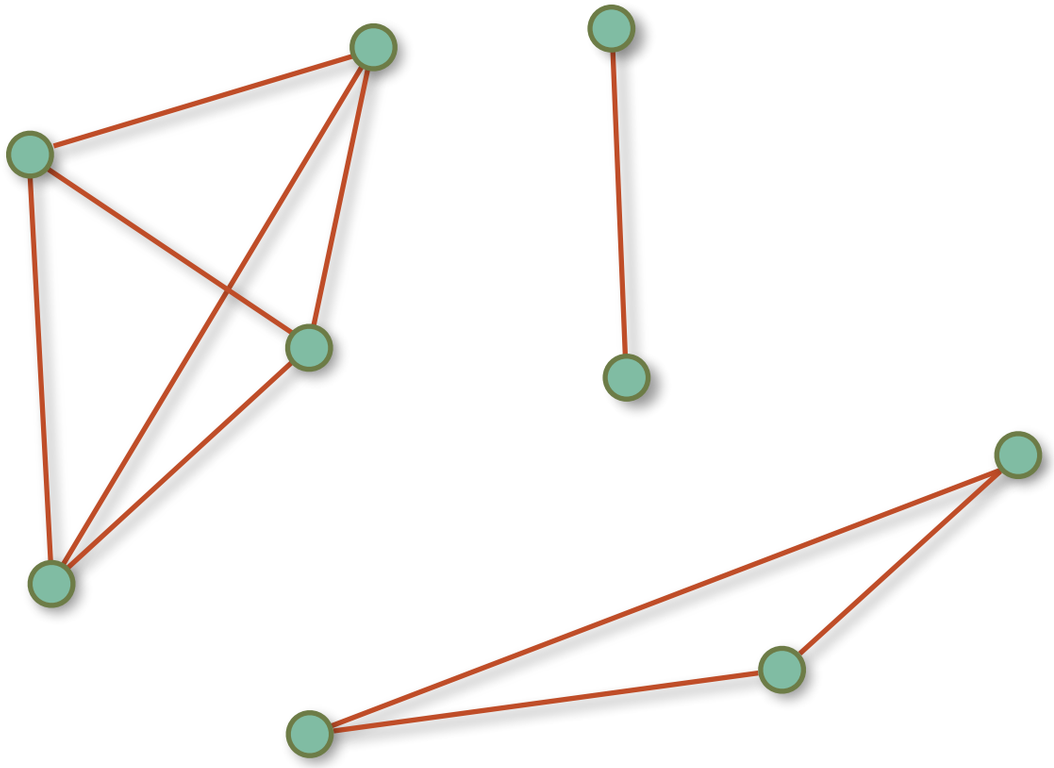
- ***Provably** solvable by block coordinate descent.*

- Interpretation as alternating between ***supervised*** and ***unsupervised*** learning steps.

Outline

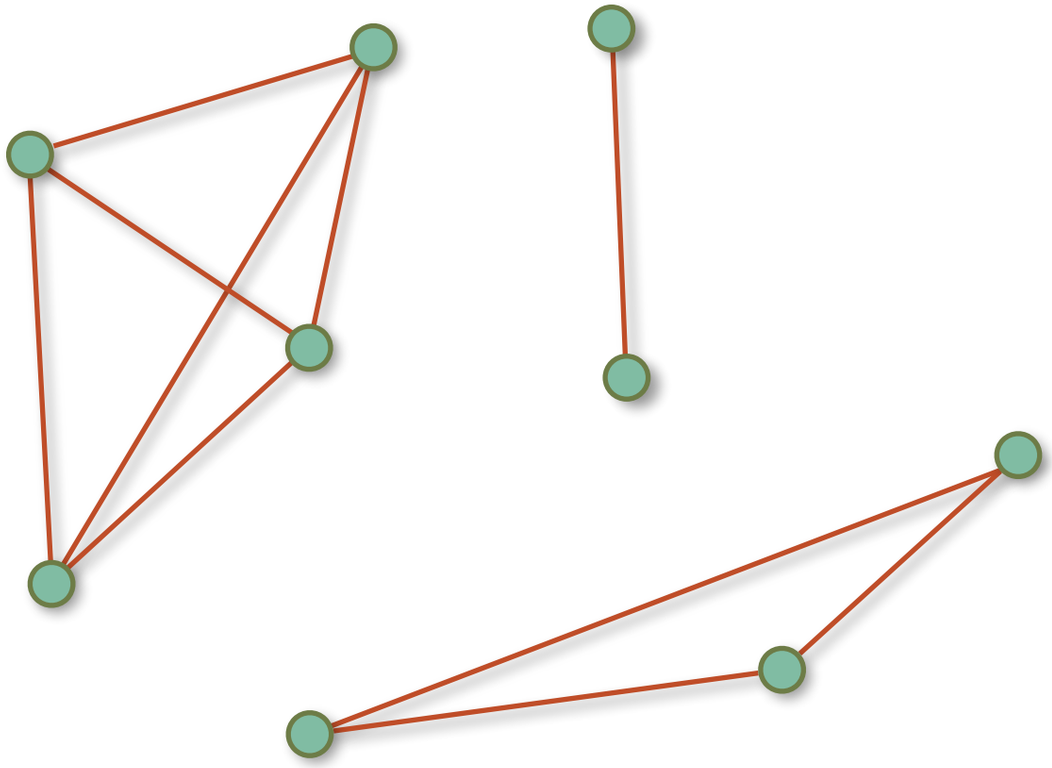
- Motivations
- Reproducing kernels Hilbert spaces for vector-valued functions
- A unifying framework for learning multiple tasks and their structure
- **Recovering previous examples from the literature**
- A novel sparse relations learning framework
- Summary





$$\underset{f \in \mathcal{H} \;\; A \in S_+^T}{minimize} \; \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

[Jacob et al. ’08]

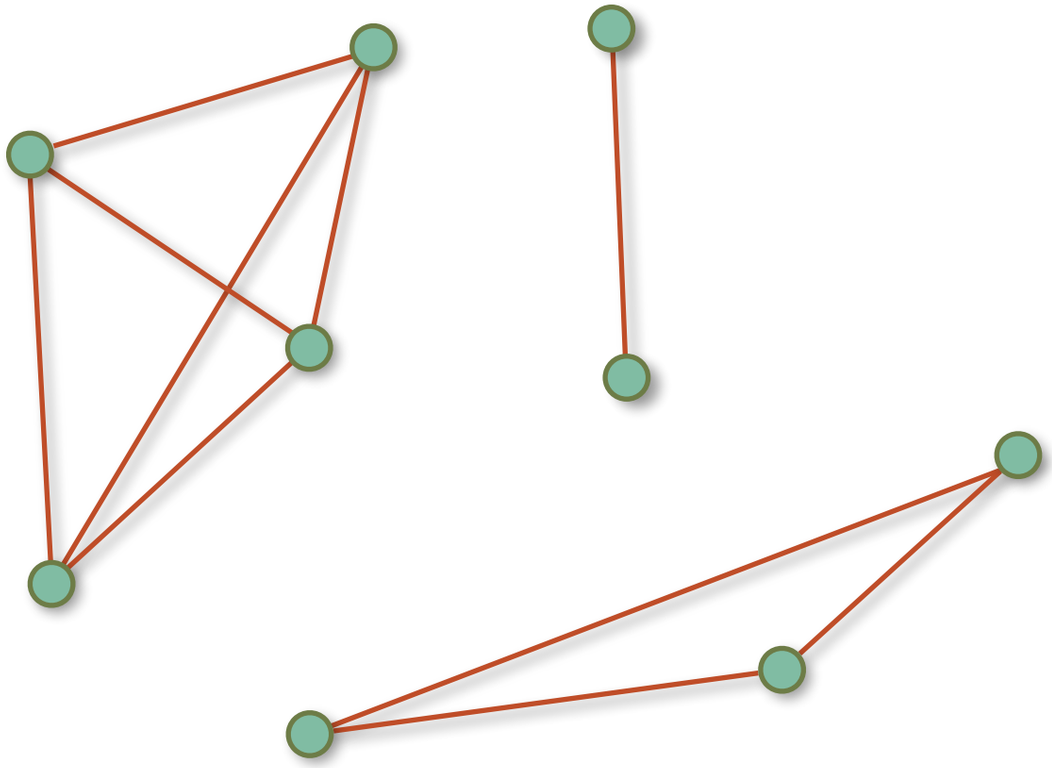


$$M = I - L$$

L Normalized Graph Laplacian

$$\underset{f \in \mathcal{H} \;\; A \in S_+^T}{minimize} \; \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

[Jacob et al. ’08]



$$M = I - L$$

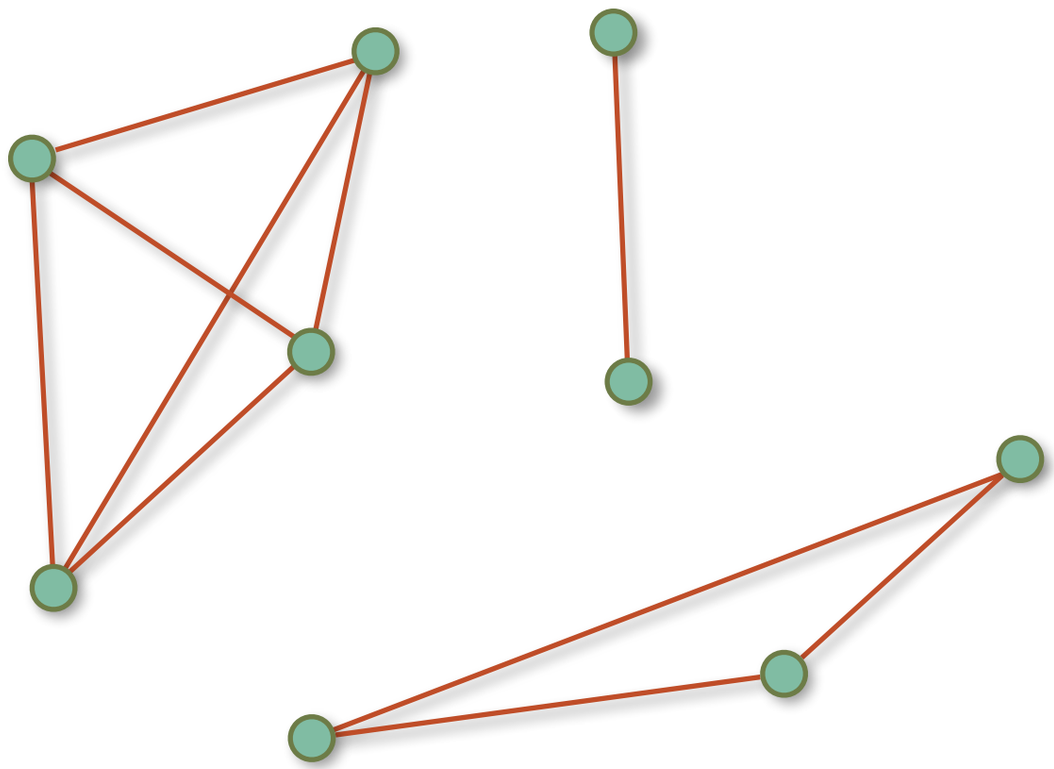
L Normalized Graph Laplacian

$$\underset{f \in \mathcal{H} \;\; A \in S_+^T}{minimize} \; \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + F(A)$$

Indicator function of the set

$$A^{-1}(M) = \epsilon_M \frac{1}{T} \mathbf{1}\mathbf{1}^\top + \epsilon_B (M - \frac{1}{T} \mathbf{1}\mathbf{1}^\top) + \epsilon_W (I - M)$$

[Jacob et al. ’08]



Mean

$$\|\bar{f}\|_k^2 = \frac{1}{T} tr(\mathbf{1}\mathbf{1}^\top B^\top K B)$$

Between-clusters Variance

$$\sum_{c=1}^{r} \|\bar{f}_c - \bar{f}\|_k^2 = tr((M - \frac{1}{T}\mathbf{1}\mathbf{1}^\top) B^\top K B)$$

Within-cluster Variance

$$\sum_{c=1}^{r} \sum_{i \in \mathcal{J}(c)} \|f_i - \bar{f}_c\|_k^2 = tr((I - M) B^\top K B)$$

$$\underset{f \in \mathcal{H}\ A \in S_+^T}{minimize} \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \|f\|_{\mathcal{H}}^2$$

$$A^{-1}(M) = \epsilon_M \frac{1}{T}\mathbf{1}\mathbf{1}^\top + \epsilon_B (M - \frac{1}{T}\mathbf{1}\mathbf{1}^\top) + \epsilon_W (I - M)$$

[Jacob et al. ’08]

Shared Feature Structure

$$A = I$$

$$\Gamma(x, x') = k(x, x')A \qquad \text{Separable}$$

$$k(x, x') = \langle x, Dx' \rangle_{\mathcal{X}} \qquad \text{Linear (scalar) Kernel}$$

Shared Feature Structure

$$\Gamma(x, x') = k(x, x')I \qquad \text{Separable}$$

$$k(x, x') = \langle x, Dx' \rangle_{\mathcal{X}} \qquad \text{Linear (scalar) Kernel}$$

Shared Feature Structure

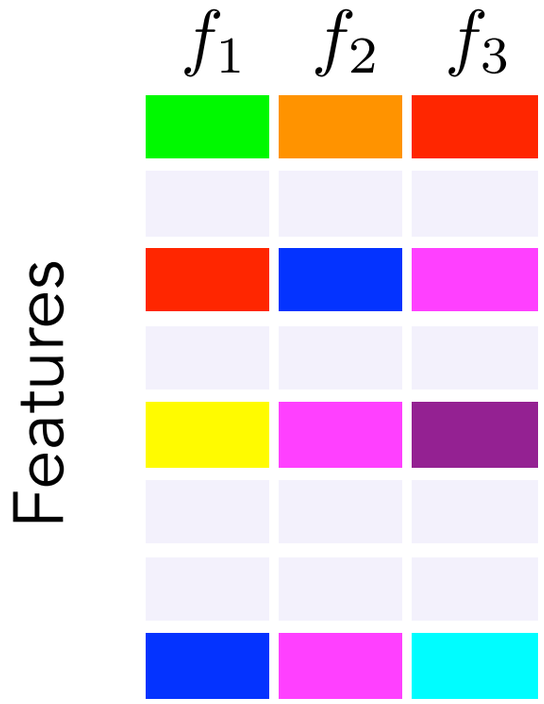
$$\Gamma(x, x') = k(x, x')I \qquad \text{Separable}$$

$$k(x, x') = \langle x, Dx' \rangle_{\mathcal{X}} \qquad \text{Linear (scalar) Kernel}$$

$$\underset{\substack{f \subset \mathcal{H} \\ D \; diagonal, D \succeq 0}}{minimize} \; \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + \lambda \; trace(D)$$

[Argyriou et al. ‘08]

Shared Feature Structure



$$\Gamma(x, x') = k(x, x')I$$ Separable

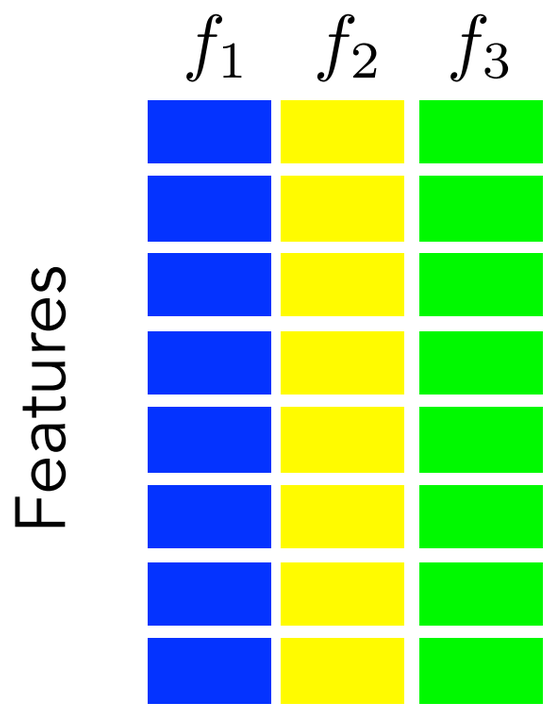
$$k(x, x') = \langle x, Dx' \rangle_{\mathcal{X}}$$ Linear (scalar) Kernel

Shared Sparsity

$$\underset{\substack{f \in \mathcal{H} \\ D\ diagonal, D \succeq 0}}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + \lambda\ trace(D)$$

[Argyriou et al. ‘08]

Shared Feature Structure



$$\Gamma(x, x') = k(x, x')I$$ Separable

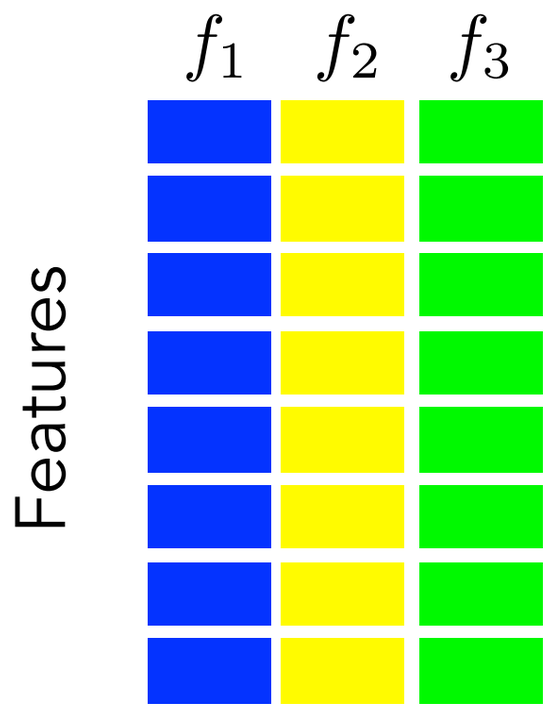
$$k(x, x') = \langle x, Dx' \rangle_{\mathcal{X}}$$ Linear (scalar) Kernel

Low-rank

$$\underset{\substack{f \in \mathcal{H} \\ D \succeq 0}}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + \lambda \; trace(D)$$

[Argyriou et al. '08]

Shared Feature Structure



$$\Gamma(x, x') = k(x, x')A$$ Separable

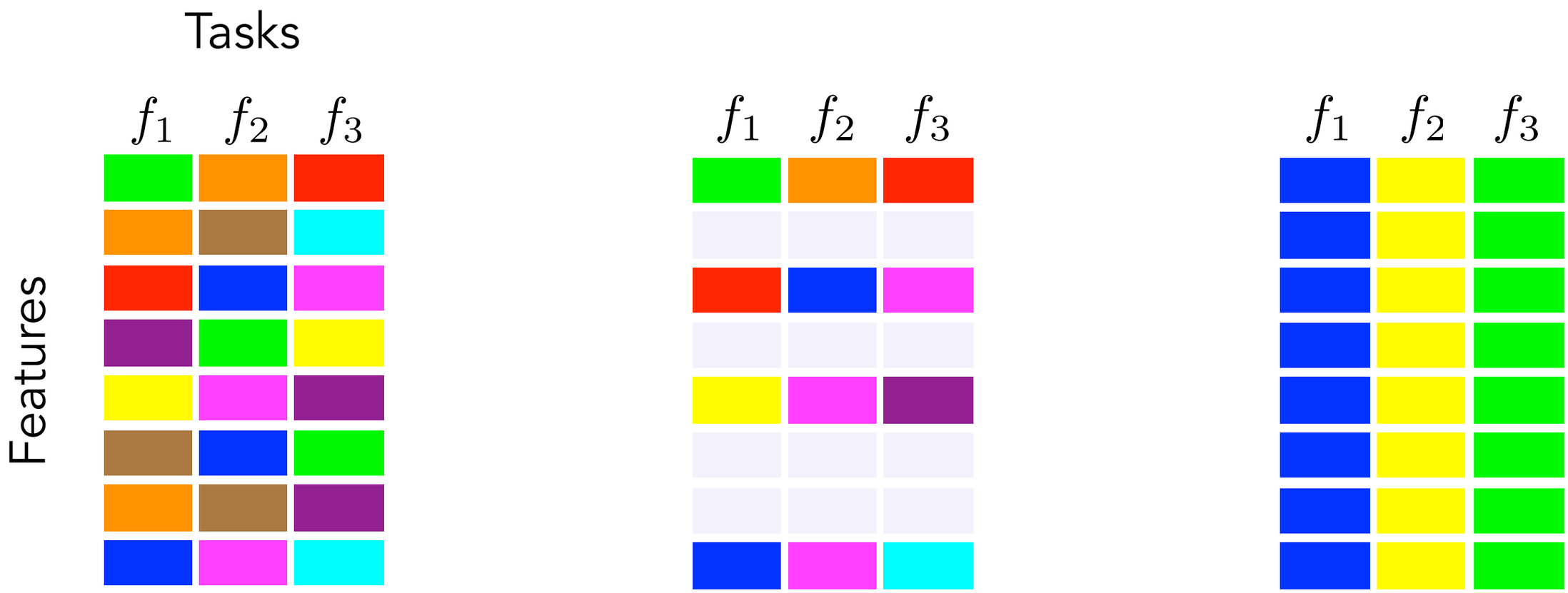
$$k(x, x') = \langle x, x' \rangle_{\mathcal{X}}$$ Linear (scalar) Kernel

Low-rank

$$\underset{\substack{f \in \mathcal{H} \\ A \succeq 0}}{minimize} \quad \frac{1}{T} \sum_{\substack{t=1 \\ i=1}}^{T, n_t} \frac{1}{n_t} \mathcal{L}(y_{it}, f_t(x_{it})) + \lambda \| f \|_{\mathcal{H}}^2 + \lambda \, trace(A)$$

[Argyriou et al. ‘08]

Shared Feature Structure



Independent *Shared sparsity* *Low-rank*

Idea: Enforce relations by imposing shared structure on the input space.

Example: Shared sparsity/group sparsity, Low-dimensional feature space…

[Argyriou et al. ‘08]

Sparse Multi-tasks Relations Learning

[Ciliberto et al. ‘13]

Sparse Multi-tasks Relations Learning

$$f_t(x) = \sum_{i=1}^{n} k(x, x_i) A_t^\top c_i = \sum_{s=1}^{T} A_{ts} g_s(x) \qquad g_s(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i) c_{is} \in \mathcal{H}_k$$

[Ciliberto et al. ‘13]

Sparse Multi-tasks Relations Learning [Ciliberto et al '15]

$$f_t(x) = \sum_{i=1}^{n} k(x, x_i) A_t^\top c_i = \sum_{s=1}^{T} A_{ts} g_s(x) \qquad g_s(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i) c_{is} \in \mathcal{H}_k$$

$$F(A) = \|A\|_{\ell_1}$$

Sparse Multi-tasks Relations Learning [Ciliberto et al '15]

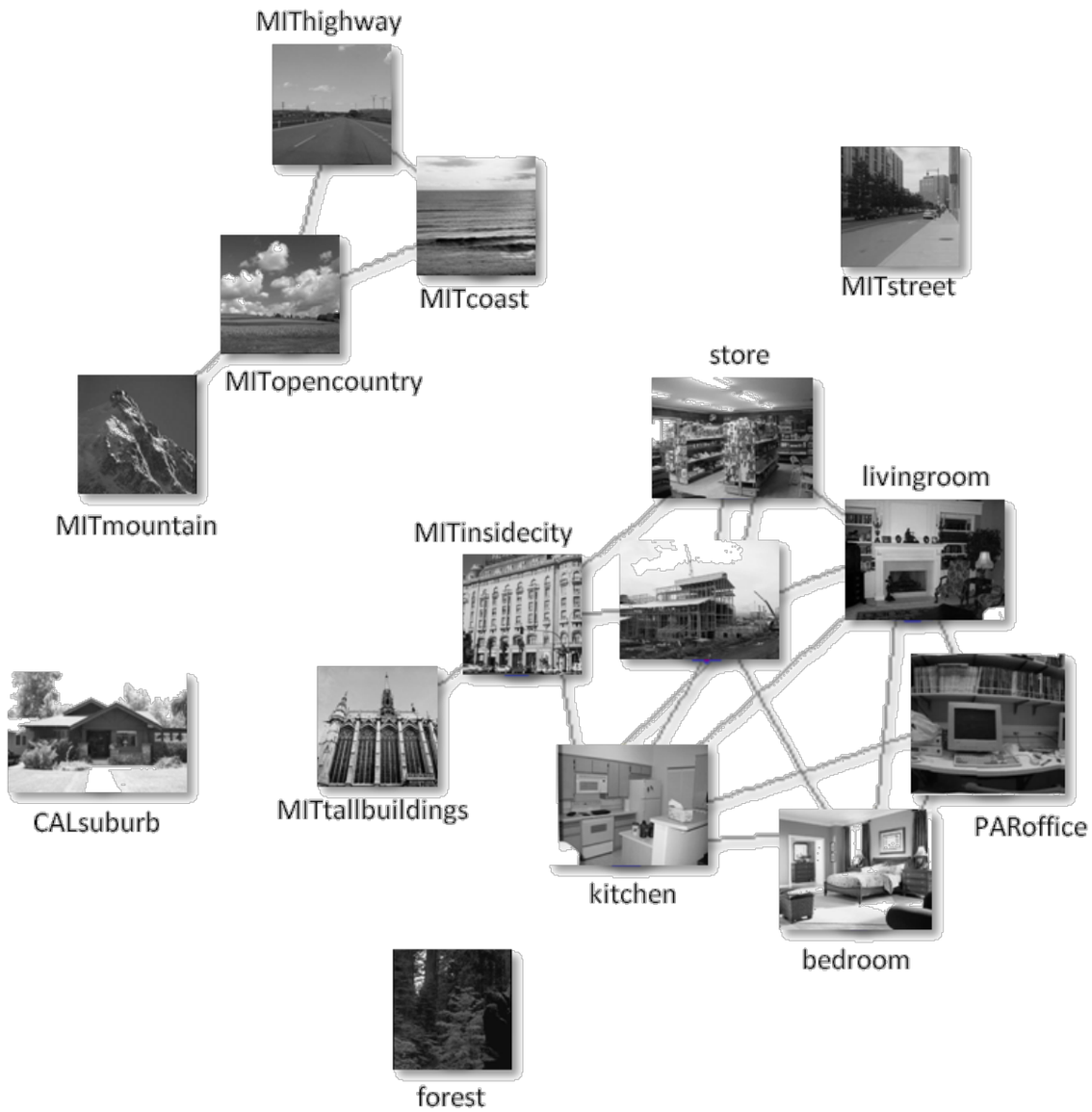
$$f_t(x) = \sum_{i=1}^{n} k(x, x_i) A_t^\top c_i = \sum_{s=1}^{T} A_{ts} g_s(x)$$

$$g_s(\cdot) = \sum_{i=1}^{n} k(\cdot, x_i) c_{is} \in \mathcal{H}_k$$

$$F(A) = \|A\|_{\ell_1}$$

	Baseline	Tr	Fro	Sparse
Acc. (%)	79.2	80.1	80.0	81.3
Var (%)	±0.01	±0.03	±0.01	±0.08

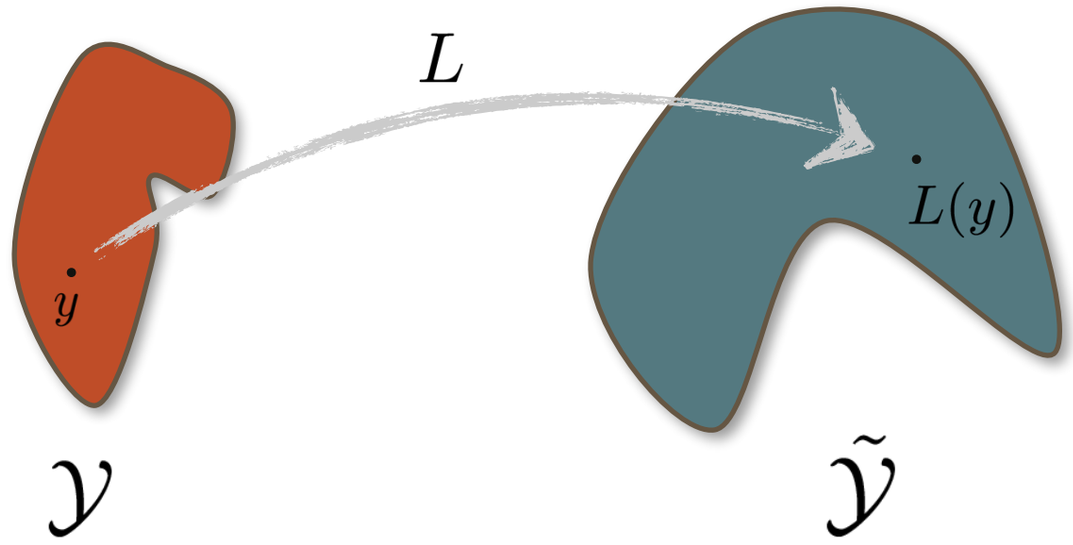
Classification Results on the
15 - Scenes Dataset



Embeddings

Idea: Embed $\mathcal{Y}$ in a new $\tilde{\mathcal{Y}}$ to account for complex structures

Example: Words, Threes, Graphs…



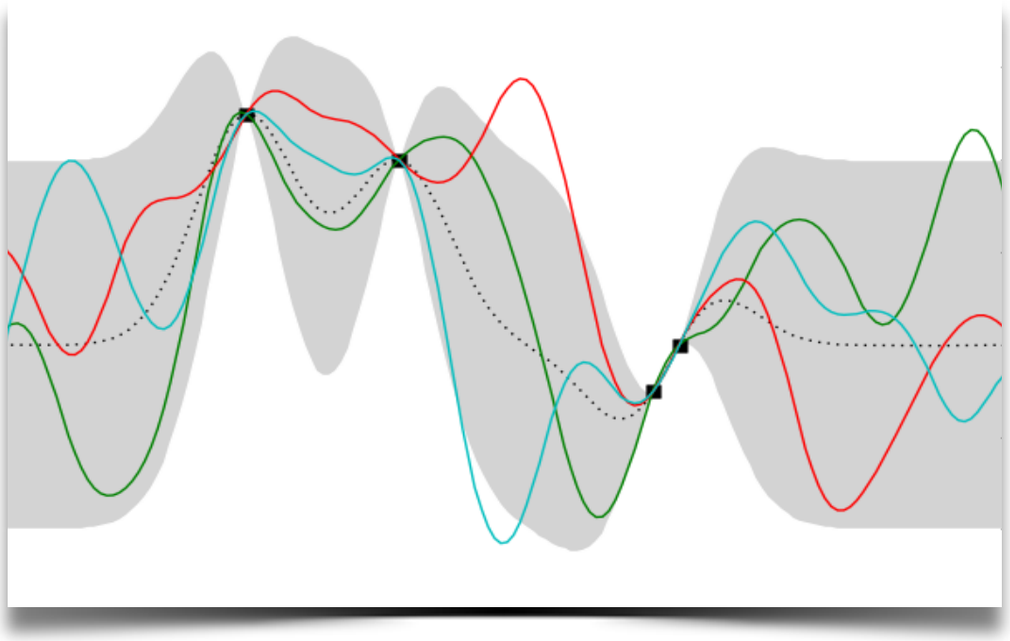
[Fergus et al. ’10, Weston et al. ’12, Joachims et al. ’09, Crammer and Singer ’00]

Gaussian Processes

$$f = (f_1, \ldots, f_T) \sim \mathcal{GP}(\mathbf{m}, \mathbf{K})$$

$\mathbf{m} : \mathcal{X} \to \mathbb{R}^T$ Bias

$\Gamma : \mathcal{X} \times \mathcal{X} \to \mathbb{R}^{T \times T}$ Covariance



[Alvarez ’12]

Summary

- A general **convex** framework that ***recovers*** several previous methods as special cases.

Summary

- A general **convex** framework that ***recovers*** several previous methods as special cases.

- A **provably convergent** algorithm for alternating minimization (and Block Coordinate Descent).

Summary

- A general **convex** framework that ***recovers*** several previous methods as special cases.

- A **provably convergent** algorithm for alternating minimization (and Block Coordinate Descent).

- A novel setting for learning **sparse relations** among tasks.

Summary

- A general **convex** framework that ***recovers*** several previous methods as special cases.
- A **provably convergent** algorithm for alternating minimization (and Block Coordinate Descent).
- A novel setting for learning **sparse relations** among tasks.

Future Directions

- (Short Term) Explore new structure-inducing penalties for e.g. **unknown** number of clusters, **sparse** conditional independence among tasks.
- (Long Term) **Beyond separable kernels**: parametrize a family of non-separable matrix-valued kernels able to encode **joint** input-output structures.

Questions?