Supervised learning so far

- **Regression** \( f : X \rightarrow Y \subseteq \mathbb{R} \)
- **Classification** \( f : X \rightarrow Y = \{-1, 1\} \)

What next?
- **Vector-valued** \( f : X \rightarrow Y \subseteq \mathbb{R}^T \)
- **Multiclass** \( f : X \rightarrow Y = \{1, 2, \ldots, T\} \)
- ...
Multitask learning

Given

\[ S_1 = (x_1^1, y_1^1)_{i=1}^{n_1}, \ldots, S_T = (x_T^T, y_T^T)_{i=1}^{n_T} \]

find

\[ f_1 : X_1 \rightarrow Y_1, \ldots, f_T : X_T \rightarrow Y_T \]
Multitask learning

Given

\[ S_1 = (x_i^1, y_i^1)_{i=1}^{n_1}, \ldots, S_T = (x_i^T, y_i^T)_{i=1}^{n_T} \]

find

\[ f_1 : X_1 \to Y_1, \ldots, f_T : X_T \to Y_T \]

▶ vector valued regression,

\[ S_n = (x_i, y_i)_{i=1}^{n}, \quad x_i \in X, \quad y_i \in \mathbb{R}^T \]

MTL with equal inputs! Output coordinates are “tasks”

▶ multiclass

\[ S_n = (x_i, y_i)_{i=1}^{n}, \quad x_i \in X, \quad y_i \in \{1, \ldots, T\} \]
Why MTL?
Why MTL?

Real data!
Why MTL?

Related problems:
- conjoint analysis
- transfer learning
- collaborative filtering
- co-kriging

Examples of applications:
- geophysics
- music recommendation (Dinuzzo 08)
- pharmacological data (Pillonetto at el. 08)
- binding data (Jacob et al. 08)
- movies recommendation (Abernethy et al. 08)
- HIV Therapy Screening (Bickel et al. 08)
Why MTL?

VVR, e.g. vector fields estimation
Why MTL?

Component 1

Component 2

L. Rosasco, RegML 2020
Penalized regularization for MTL

$$\text{err}(w_1, \ldots, w_T) + \text{pen}(w_1, \ldots, w_T)$$

We start with linear models

$$f_1(x) = w_1^\top x, \ldots, f_T(x) = w_T^\top x$$
Empirical error

\[ \hat{\mathcal{E}}(w_1, \ldots, w_T) = \sum_{i=1}^{T} \frac{1}{n_i} \sum_{j=1}^{n_i} (y_{ij} - w_i^T x_{ij})^2 \]

- could consider other losses
- could try to "couple" errors
Least squares error

We focus on vector valued regression (VVR)

\[ S_n = (x_i, y_i)_{i=1}^n, \quad x_i \in X, \quad y_i \in \mathbb{R}^T \]
Least squares error

We focus on vector valued regression (VVR)

\[ S_n = (x_i, y_i)_{i=1}^n, \quad x_i \in X, \quad y_i \in \mathbb{R}^T \]

\[
\frac{1}{n} \sum_{t=1}^T \sum_{i=1}^n (y_{ti} - w_t^\top x_i)^2 = \frac{1}{n} \| \hat{X} W - \hat{Y} \|^2_F
\]

\[
\| W \|^2_F = \text{Tr}(W^\top W), \quad W = (w_1, \ldots, w_T), \quad \hat{Y}_{it} = \hat{y}_{ti} \quad i = 1 \ldots n \quad t = 1 \ldots T
\]
MTL by regularization

pen(\(w_1 \ldots w_T\))

- Coupling task solutions by regularization
- Borrowing strength
- Exploit structure
Regularizations for MTL

\[
\text{pen}(w_1, \ldots, w_T) = \sum_{t=1}^{T} \|w_t\|^2
\]
Regularizations for MTL

\[ \text{pen}(w_1, \ldots, w_T) = \sum_{t=1}^{T} \|w_t\|^2 \]

Single tasks regularization!

\[
\min_{w_1, \ldots, w_T} \frac{1}{n} \sum_{t=1}^{T} \sum_{i=1}^{n} (y_i^t - w_t^\top x_i)^2 + \lambda \sum_{t=1}^{T} \|w_t\|^2 = \\
\sum_{t=1}^{T} \left( \min_{w_t} \frac{1}{n} \sum_{i=1}^{n} (y_i^t - w_t^\top x_i)^2 + \lambda \|w_t\|^2 \right)
\]
Isotropic coupling

\[(1 - \alpha) \sum_{j=1}^{T} \|w_j\|^2 + \alpha \sum_{j=1}^{T} \left\| w_j - \frac{1}{T} \sum_{i=1}^{T} w_i \right\|^2 \]
Regularizations for MTL

▶ Isotropic coupling

$$(1 - \alpha) \sum_{j=1}^{T} \| w_j \|^2 + \alpha \sum_{j=1}^{T} \left\| w_j - \frac{1}{T} \sum_{i=1}^{T} w_i \right\|^2$$

▶ Graph coupling - Let $M \in \mathbb{R}^{T \times T}$ an adjacency matrix, with $M_{ts} \geq 0$

$$\sum_{t=1}^{T} \sum_{s=1}^{T} M_{ts} \| w_t - w_s \|^2 + \gamma \sum_{t=1}^{T} \| w_t \|^2$$

special case: output divided in clusters
A general form of regularization

All the regularizers so far are of the form

$$\sum_{t=1}^{T} \sum_{s=1}^{T} A_{ts} w_t^\top w_s$$

for a suitable positive definite matrix $A$.
MTL regularization revisited

▶ Single tasks $\sum_{j=1}^{T} \|w_j\|^2 \implies A = I$

$\text{L.Rosasco, RegML 2020}$
MTL regularization revisited

- Single tasks $\sum_{j=1}^{T} \| w_j \|^2 \implies A = I$
- Isotropic coupling

$$(1 - \alpha) \sum_{j=1}^{T} \| w_j \|^2 + \alpha \sum_{j=1}^{T} \left\| w_j - \frac{1}{T} \sum_{j=1}^{T} w_j \right\|^2$$

$\implies A = I - \frac{\alpha}{T} \mathbf{1}$$
MTL regularization revisited

- Single tasks \( \sum_{j=1}^{T} \|w_j\|^2 \implies A = I \)

- Isotropic coupling

\[
(1 - \alpha) \sum_{j=1}^{T} \|w_j\|^2 + \alpha \sum_{j=1}^{T} \left\| w_j - \frac{1}{T} \sum_{j=1}^{T} w_j \right\|^2 \implies A = I - \frac{\alpha}{T} \mathbf{1}
\]

- Graph coupling

\[
\sum_{t=1}^{T} \sum_{s=1}^{T} M_{ts} \|w_t - w_s\|^2 + \gamma \sum_{t=1}^{T} \|w_t\|^2 \implies A = L + \gamma I,
\]

where \( L \) graph Laplacian of \( M \)

\[
L = D - M, \quad D = \text{diag}(\sum_{j} M_{1,j}, \ldots, \sum_{j} M_{T,j})
\]

L. Rosasco, RegML 2020
A general form of regularization

Let $W = (w_1, \ldots, w_T)$, $A \in \mathbb{R}^{T \times T}$

Note that

$$\sum_{t=1}^{T} \sum_{s=1}^{T} A_{ts} w_t^\top w_s = \text{Tr}(WAW^\top)$$
A general form of regularization

Let $W = (w_1, \ldots, w_T)$, $A \in \mathbb{R}^{T \times T}$

Note that

$$
\sum_{t=1}^{T} \sum_{s=1}^{T} A_{ts} w_t^\top w_s = \text{Tr}(WAW^\top)
$$

Indeed

$$
\text{Tr}(WAW^\top) = \sum_{i=1}^{d} W_i^\top AW_i = \sum_{i=1}^{d} \sum_{t,s=1}^{T} A_{ts} W_{it} W_{is}
$$

$$
= \sum_{t,s=1}^{T} A_{ts} \sum_{i=1}^{d} W_{is} W_{ir} = \sum_{t,s=1}^{T} A_{ts} w_t^\top w_s
$$
Computations

\[ \frac{1}{n} \| \hat{X}W - \hat{Y} \|_F^2 + \lambda \text{Tr}(WAW^\top) \]
Consider the SVD $A = U\Sigma U^\top$, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_T)$.
Consider the SVD $A = U \Sigma U^\top$, $\Sigma = \text{diag}(\sigma_1, \ldots, \sigma_T)$ let

$$\tilde{W} = WU, \quad \tilde{Y} = \hat{Y}U$$

then we can rewrite the above problem as

$$\frac{1}{n} \| \hat{X}W - \hat{Y} \|^2_F + \lambda \text{Tr}(WAW^\top)$$
Finally, rewrite
\[ \frac{1}{n} \| \hat{X} \tilde{W} - \tilde{Y} \|_F^2 + \lambda \text{Tr}(\tilde{W} \Sigma \tilde{W}^\top) \]
as
\[ \sum_{t=1}^T \left( \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i^t - \tilde{w}_t^\top x_i)^2 + \lambda \sigma_t \| \tilde{w}_t \|^2 \right) \]
and use \( W = \tilde{W} U^\top \)

Compare to single task regularization...
Computations (cont.)

\[
E_\lambda(W) = \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2_F + \lambda \text{Tr}(WAW^\top)
\]

Alternatively

\[
\nabla E_\lambda(W) = 2 \frac{1}{n} \hat{X}^\top (\hat{X}W - \hat{Y}) + 2\lambda WA
\]

\[
W_{t+1} = W_t - \gamma \nabla E_\lambda(W_t)
\]

Trivially extends to other loss functions.
Beyond Linearity

\[ f_t(x) = w_t^\top \Phi(x), \quad \Phi(x) = (\phi_1(x), \ldots, \phi_p(x)) \]

\[ E_\lambda(W) = \frac{1}{n} \| \hat{\Phi}W - \hat{Y} \|^2 + \lambda \textrm{Tr}(WAW^\top), \]

with \( \hat{\Phi} \) matrix with rows \( \Phi(x_1), \ldots, \Phi(x_n) \)

L. Rosasco, RegML 2020
Nonparametrics and kernels

\[ f_t(x) = \sum_{i=1}^{n} K(x, x_i) C_{it} \]

with

\[ C_{\ell+1} = C_\ell - \gamma \left( \frac{2}{n} \hat{K} C_\ell - \hat{Y} + 2\lambda C_\ell A \right) \]

- \( C_\ell \in \mathbb{R}^{n \times T} \)
- \( \hat{K} \in \mathbb{R}^{n \times n}, \hat{K}_{ij} = K(x_i, x_j) \)
- \( \hat{Y} \in \mathbb{R}^{n \times T}, \hat{Y}_{ij} = y_i^j \)
Spectral filtering for MTL

Beyond penalization

$$\min_W \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WAW^\top),$$

other forms of regularizations can be considered

- projection
- early stopping
Multiclass and MTL

\[ Y = \{1, \ldots, T\} \]
From Multiclass to MTL

Encoding For $j = 1, \ldots, T$

$$j \mapsto e_j$$

canonical vector of $\mathbb{R}^T$

the problem reduces to vector valued regression

Decoding For $f(x) \in \mathbb{R}^T$

$$f(x) \mapsto \arg\max_{t=1,\ldots,T} e_t^T f(x) = \arg\max_{t=1,\ldots,T} f_t(x)$$

L. Rosasco, RegML 2020
Single MTL and OVA

Write

$$\min_W \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WW^\top),$$

as

$$\sum_{t=1}^T \min_{w_t} \frac{1}{n} \sum_{i=1}^{n_t} (w_t^\top x_i - y_i)^2 + \lambda \|w_t\|^2$$

This is known as one versus all (OVA)
Beyond OVA

Consider

\[
\min_W \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WAW^\top),
\]

that is

\[
\sum_{t=1}^T \min_{\tilde{w}_t} \sum_{t=1}^T \left( \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i^t - \tilde{w}_t^\top x_i)^2 + \lambda \sigma_t \| \tilde{w}_t \|^2 \right)
\]

Class relatedness encoded in \( A \)
Back to MTL

\[
\sum_{t=1}^{T} \frac{1}{n_t} \sum_{j=1}^{n_t} (y^t_j - w^\top_i x^t_j)^2
\]

\[
\Downarrow
\]

\[
\| (\hat{X} W - \overline{Y}) \odot M \|_F^2, \quad n = \sum_{t=1}^{T} n_t
\]

- \textcircled{\odot} Hadamard product
- \(M\) mask
- \(\overline{Y}\) having one non-zero value for each row
Computations

$$\min_W \|(\hat{X}W - \bar{Y}) \odot M\|_F^2 + \lambda \text{Tr}(WAW^\top)$$

- can be rewritten using tensor calculus
- computation for vector valued regression easily extended
- sparsity of $M$ can be exploited
From MTL to matrix completion

Special case Take $d = n$ and $X = I$

$$\| (\hat{X}W - \overline{Y}) \circ M \|_F^2$$

$$\downarrow$$

$$\sum_{t=1}^{T} \sum_{i=1}^{n} (w_{ij} - \bar{y}_{ij})^2 M_{ij}$$
Summary so far

A regularization framework for
- VVR
- Multiclass
- MTL
- Matrix completion

if the structure of the “tasks” is known.

What if it is not?
The structure of MTL

Consider

$$\min_{W} \frac{1}{n} \| \hat{X} W - \hat{Y} \|^2 + \lambda \text{Tr}(WAW^\top),$$

the matrix $A$ encodes structure.

Can we learn it?
Learning structure of MTL

Consider

$$\min_{W,A} \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WAW^\top) + \gamma \text{pen}(A)$$

Estimate a positive definite matrix $A$ using a regularizer $\text{pen}(A)$
Regularizers for MTL

For example consider

$$\min_{W,A} \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WAW^\top) + \gamma \text{Tr}(A^{-2})$$

using the same change of coordinates as before we have

$$\min_{\tilde{w}_1,\ldots,\tilde{w}_T, \sigma_1,\ldots,\sigma_T} \sum_{t=1}^T \sum_{t=1}^T \frac{1}{n} \sum_{i=1}^n (\tilde{y}_i^t - \tilde{w}_t^\top x_i)^2 + \lambda \sigma_t \| \tilde{w}_t \|^2 + \gamma \sum_{t=1}^T \left( \frac{1}{\sigma_t^2} \right)$$

we avoid each task having too little weight
Alternating minimization

Solving

$$\min_{W,A} \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WAW^\top) + \gamma \text{pen}(A)$$
Alternating minimization

Solving

\[
\min_{W,A} \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WA_0W^\top) + \gamma \text{pen}(A)
\]

- Fix \( A = A_0 \)
- Compute \( W_1 \) solving

\[
\min_{W} \frac{1}{n} \| \hat{X}W - \hat{Y} \|^2 + \lambda \text{Tr}(WA_0W^\top)
\]

- Compute \( A_1 \) solving

\[
\min_{A} \lambda \text{Tr}(W_1AW_1^\top) + \gamma \text{pen}(A)
\]

- Repeat…
This class

- Why MTL?
- Regularization for MTL to exploit structure
- MTL and other problems
- Learning tasks AND their structure
Next class

Sparsity!