Consistent Multitask Learning with Nonlinear Output Constraints

Carlo Ciliberto
Department of Computer Science, UCL

joint work w/ Alessandro Rudi, Lorenzo Rosasco and Massi Pontil
Multitask Learning (MTL)

MTL Mantra:
leverage the similarities among multiple learning problems (tasks) to reduce the complexity of the overall learning process.

Prev. Literature:
investigated linear tasks relations (more on this in a minute).

This work:
we address the problem of learning multiple tasks that are nonlinearly related one to the other
MTL Setting

Given \( T \) datasets \( S_t = (x_{it}, y_{it})_{i=1}^{n_t} \) learn \( \hat{f}_t : \mathcal{X} \to \mathbb{R} \) by solving

\[
(\hat{f}_1, \ldots, \hat{f}_T) = \text{argmin}_{f_1, \ldots, f_T \in \mathcal{H}} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(f_t, S_t) + R(f_1, \ldots, f_T)
\]

- \( \mathcal{H} \) space of hypotheses.
- \( \mathcal{L}(f_t, S_t) = \frac{1}{n_t} \sum_{i=1}^{n_t} \ell(f_t(x_{it}, y_{it})) \) Data fitting term. Loss \( \ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R} \) (e.g. least squares, logistic, hinge, etc.).
- \( R(f_1, \ldots, f_T) \) a joint tasks-structure regularizer
Previous Work: Linear MTL

For example $R(f_1, \ldots, f_T) =$

- Single task learning
  \[ \lambda \sum_{t=1}^{T} \| f_t \|^2_{\mathcal{H}} \]

- Variance Regularization
  \[ \lambda \sum_{t=1}^{T} \| f_t - \bar{f} \|^2_{\mathcal{H}} \text{ with } \bar{f} = \frac{1}{T} \sum_{t=1}^{T} \]

- Clustered tasks
  \[ \lambda_1 \sum_{t \in C(c)} \| f_t - \bar{f}_c \|^2_{\mathcal{H}} + \lambda_2 \sum_{c=1}^{\left| C \right|} \| \bar{f}_c - \bar{f} \|^2_{\mathcal{H}} \]

- Similarity regularizer
  \[ \lambda \sum_{t,s} W_{s,t} \| f_t - f_s \|^2_{\mathcal{H}} \quad W_{s,t} \geq 0 \]

Why “Linear”? Because the tasks relations are encoded in a matrix.

\[ R(f_1, \ldots, f_T) = \sum_{t,s=1}^{T} A_{t,s} \langle f_t, f_s \rangle_{\mathcal{H}} \quad \text{with} \quad A \in \mathbb{R}^{T \times T} \]
Nonlinear MTL: Setting

What if relations are nonlinear? We study the case where tasks satisfy a set of $k$ equations $\gamma(f_1(x), \cdots, f_T(x)) = 0$ identified by $\gamma : \mathbb{R}^T \to \mathbb{R}^k$.

Examples

- Manifold-valued learning
- Physical systems (e.g. robotics)
- Logical constraints (e.g. ranking)
**Nonlinear MTL: Setting**

**NL-MTL Goal:** approximate $f^* : \mathcal{X} \to \mathcal{C}$ minimizer the **Expected Risk**

$$
\min_{f : \mathcal{X} \to \mathcal{C}} \mathcal{E}(f), \quad \mathcal{E}(f) = \frac{1}{T} \int \ell(f_t(x), y) \, d\rho_t(x, y)
$$

where

- $f : \mathcal{X} \to \mathcal{C}$ is such that $f(x) = (f_1(x), \ldots, f_T(x))$ for all $x \in \mathcal{X}$.
- $\mathcal{C} = \{ c \in \mathbb{R}^T | \gamma(c) = 0 \}$ is the **constraints set** induced by $\gamma$.
- $\rho_t(x, y) = \rho_t(y|x)\rho_X(x)$ is the **unknown** data distribution.
Nonlinear MTL: Challenges

Why not try **Empirical Risk Minimization**?

\[ \hat{f} = \arg\min_{\mathcal{H}\subset\{f: \mathcal{X}\rightarrow\mathcal{C}\}} \frac{1}{T} \sum_{t=1}^{T} \mathcal{L}(f_t, S_t) \]

Problems:

- **Modeling**: \( f_1, f_2 : \mathcal{X} \rightarrow \mathcal{C} \) does not guarantee \( f_1 + f_2 : \mathcal{X} \rightarrow \mathcal{C} \). \( \mathcal{H} \) not a linear space. How to choose a “good” \( \mathcal{H} \) in practice?

- **Computations**: Hard (non-convex) optimization. How to solve it?

- **Statistics**: How to study the generalization properties of \( \hat{f} \)?
Idea: formulate NL-MTL as a *structured prediction* problem.

\[ \mathcal{X} \xrightarrow{f} \mathcal{C} \]

*Structured Prediction*: originally designed for discrete outputs, but recently generalized to any set \( \mathcal{C} \) within the **SELF** framework [Ciliberto et al. 2016].
We propose to approximate $f^*$ via the estimator $\hat{f} : \mathcal{X} \rightarrow \mathcal{C}$ such that

$$\hat{f}(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \alpha_{it}(x) \ell(c_t, y_{it})$$

where the weights are obtained in closed form as

$$(\alpha_{i1}(x), \cdots, \alpha_{in_t}(x)) = (K_t + \lambda I)^{-1} v_t(x)$$

with $K_t$ the kernel matrix $(K_t)_{ij} = k(x_{it}, x_{jt})$ of $t$-th dataset and $v_t(x) \in \mathbb{R}^n$ with $v_t(x)_i = k(x_{it}, x)$. $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$ a kernel.

**Note.** evaluating $\hat{f}(x)$ requires solving an optimization over $\mathcal{C}$ (e.g. for $\ell$ least squares it $\hat{f}$ reduces to a projection onto $\mathcal{C}$).
Theoretical Results

Thm. 1 (Universal Consistency)

\[ \mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \to 0 \quad \text{with probability } 1. \]

Thm. 2 (Rates). Let \( n = n_t \) and \( g^*_t \in \mathcal{G} \) for all \( t = 1, \ldots, T \). Then

\[ \mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq O(n^{-1/4}) \quad \text{with high probability} \]

Thm. 3 (Benefits of MTL). Let \( C \subset \mathbb{R}^T \) radius 1 sphere. Let \( N = nT \).

Then \[ \mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq O(N^{-1/2}) \quad \text{with high probability} \]
Intuition

Ok... but how did we get there?
Def. $\ell : \mathcal{C} \times \mathcal{Y} \rightarrow \mathbb{R}$ is a *structure encoding loss function* (SELF) if there exist $\mathcal{H}$ Hilbert space and $\psi : \mathcal{C} \rightarrow \mathcal{H}$, $\varphi : \mathcal{Y} \rightarrow \mathcal{H}$ such that

$$\ell(c, y) = \langle \psi(c), \varphi(y) \rangle_{\mathcal{H}} \quad \forall c \in \mathcal{C}, \; \forall y \in \mathcal{Y}.$$  

Abstract definition... BUT “most” loss functions used in MTL settings are SELF! More precisely any Lipschitz continuous function differentiable almost everywhere (e.g. least squares, logistic, hinge).
Nonlinear MTL + SELF

Minimizer of the expected risk

\[ f^*(x) = \arg\min_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \int \ell(c_t, y) \rho_t(y|x) \]
Nonlinear MTL + SELF

Minimizer of the expected risk

\[ f^*(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \int \langle \psi(c_t), \varphi(y) \rangle \mathcal{H} \rho_t(y|x) \]
Minimizer of the expected risk

\[ f^*(x) = \arg \min_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \left\langle \psi(c_t), \int \varphi(y) \rho_t(y|x) \right\rangle_H \]
Minimizer of the expected risk

$$f^*(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), g_t^*(x) \rangle_{\mathcal{H}}$$

where $g_t^* : \mathcal{X} \rightarrow \mathcal{H}$ is such that $g_t^*(x) = \int \varphi(y) \rho_t(y|x)$.
Nonlinear MTL Estimator

Idea, learn a $\hat{g}_t : \mathcal{X} \rightarrow \mathcal{H}$ for each $g^*_t$. Then approximate

$$f^*(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), g^*_t(x) \rangle_{\mathcal{H}}$$

with $\hat{f} : \mathcal{X} \rightarrow \mathcal{C}$

$$\hat{f}(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), \hat{g}_t(x) \rangle_{\mathcal{H}}$$
Nonlinear MTL Estimator

**This work:** learn $\hat{g}_t$ via kernel ridge regression. Let $G^1$ be a reproducing kernel Hilbert space with kernel $k : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$.

$$\hat{g}_t = \arg\min_{g \in G} \frac{1}{n_t} \sum_{i=1}^{n_t} \|g(x_{it}) - \varphi(y_{it})\|_H^2 + \lambda \|g\|_G^2$$

Then

$$\hat{g}_t(x) = \sum_{i=1}^{n_t} \alpha_{it}(x) \varphi(y_{it}) \quad (\alpha_{i1}(x), \ldots, \alpha_{in_t}(x)) = (K_t + \lambda I)^{-1} v_t(x)$$

where $K_t$ kernel matrix of $t$-th dataset, $v_t(x) \in \mathbb{R}^n$ evaluation vector $v_t(x)_i = k(x_{it}, x)$.

---

$^1$actually $G \otimes H$
Nonlinear MTL Estimator

Plugging into

$$
\hat{f}(x) = \arg\min_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), \hat{g}_t(x) \rangle
$$

by the SELF property we have

$$
\hat{f}(x) = \arg\min_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \alpha_{it}(x) \left\langle \psi(c_t), \sum_{i=1}^{n_t} \alpha_{it}(x) \varphi(y_{it}) \right\rangle
$$
Plugging into

$$\hat{f}(x) = \arg\min_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), \hat{g}_t(x) \rangle_H$$

by the SELF property we have

$$\hat{f}(x) = \arg\min_{c \in C} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \alpha_{it}(x) \langle \psi(c_t), \varphi(y_{it}) \rangle_H$$
Nonlinear MTL Estimator

Plugging into

$$\hat{f}(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), \hat{g}_t(x) \rangle_{\mathcal{H}}$$

by the SELF property we have

$$\hat{f}(x) = \arg\min_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n_t} \alpha_{it}(x) \ell(c_t, y_{it})$$

as desired.

Note that evaluating $\hat{f}(x)$ Does not require knowledge of $\mathcal{H}$, $\varphi$ or $\psi$!
Empirical Results

Synthetic data

Inverse dynamics (Sarcos)

Logic constraints (Ranking Movielens100k)