Consistent Multitask Learning with Nonlinear Output Constraints

Massimiliano Pontil

Istituto Italiano di Tecnologia and University College London

Joint work with *Carlo Ciliberto, Lorenzo Rosasco, and Alessandro Rudi*

Meeting in Mathematical Statistics, Luminy, December 18 - 22, 2017

Plan

- Problem
- Method
- Analysis
- Experiments

Multitask Learning (MTL)

Aim is to exploit similarities among multiple learning tasks in order to improve learning

$$(\hat{f}_1, \ldots, \hat{f}_T) = \underset{f_1,\ldots,f_T}{\operatorname{argmin}} \; \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} \ell\big(f_t(x_{ti}), y_{ti}\big) + \lambda R\big(f_1, \ldots, f_T\big)$$

- $S_t = (x_{ti}, y_{ti})_{i=1}^n$: i.i.d. sample from[1] a prescribed probability measure ρ_t on $\mathcal{X} \times \mathbb{R}$
- $\ell : \mathbb{R} \times \mathbb{R} \to \mathbb{R}$: loss function[2]
- R: penalty function encouraging commonalities between the tasks

[1] For simplicity we use the same sample size per task.

[2] The loss function could also depend on t.

Previous Work

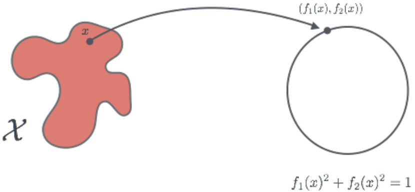
Different examples of regularizer R:

- Independent task learning: $\sum_{t=1}^{T} \|f_t\|^2$
- Similarity regularizer: $\sum_{s,t=1}^{T} A_{s,t}\|f_t - f_s\|^2 \quad A_{s,t} \geq 0$
- Groups lasso: $f_t(x) = \langle w_t, \varphi(x) \rangle, \quad \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{tj}^2}$
- Spectral regularization: $f_t(x) = \langle w_t, \varphi(x) \rangle, \quad \|\sigma([w_1 \cdots w_T])\|_1$

The above methods do not constrain the values of f, rather they encourage certain low complexity functions within a linear space.

Nonlinear MTL

We assume that f takes values on a set $\mathcal{C} \subset \mathbb{R}^T$, e.g. we prescribe a mapping $\gamma : \mathbb{R}^T \to \mathbb{R}^m$ and set $\mathcal{C} = \{y \in \mathbb{R}^T : \gamma(y_1, \ldots, y_T) = 0\}$



Examples:

- Manifold-valued learning
- Physical systems (e.g. robotics)
- Logical constraints (e.g. ranking)

Nonlinear MTL (cont.)

Goal: estimate $f^* : \mathcal{X} \to \mathcal{C}$, minimizer of the **expected risk**

$$\min_{f:\mathcal{X}\to\mathcal{C}} \mathcal{E}(f), \qquad \mathcal{E}(f) = \frac{1}{T}\sum_{t=1}^{T} \int \ell(f_t(x), y_t) \; d\rho_t(y_t, x)$$

where $f = (f_1, \dots, f_T)$ and we require that $f(x) \in \mathcal{C}, \; \forall \; x \in \mathcal{X}$

Difficulties of Empirical Risk Minimization

$$\hat{f} = \operatorname*{argmin}_{f:\mathcal{X}\to\mathcal{C}} \frac{1}{T}\sum_{t=1}^{T}\frac{1}{n}\sum_{i=1}^{n}\ell\big(f_t(x_{ti}), y_{ti}\big)$$

Problems:

- **Modeling**: $f_1, f_2 : \mathcal{X} \to \mathcal{C}$ does not guarantee $f_1 + f_2 : \mathcal{X} \to \mathcal{C}$
- **Computations**: Hard (non-convex) optimization!
- **Statistics**: How to study the generalization properties of $\hat{f}$?

We take a different path, building on [Ciliberto, Rudi, Rosasco, 2016] who considered a general structure prediction setting, showing how to reduce this problem to a simpler vector-valued learning problem

Loss Function

- **Assumption.** There exist continuous mappings $\psi : \mathbb{R} \to \mathcal{H}$ and $\phi : \mathbb{R} \to \mathcal{H}$, with $\mathcal{H}$ a Hilbert space, such that

$$\ell(y, y') = \langle \psi(y), \phi(y') \rangle \qquad \forall y, y' \in \mathbb{R}$$

- Mild assumption: verified if ℓ has derivative Lipschitz continuous almost everywhere

- Example (square loss): $\mathcal{H} = \mathbb{R}^3$, $(y - y')^2 = \big\langle (1, y, y^2), (y'^2, -2y', 1) \big\rangle$

Implication: Decomposition of the Risk

$$\begin{aligned} \mathcal{E}(f) &= \frac{1}{T}\sum_{t=1}^{T}\int \ell(f_t(x), y_t)\; d\rho_t(y_t, x) \\ &= \frac{1}{T}\sum_{t=1}^{T}\int \langle \psi(f_t(x)), \phi(y_t)\rangle d\rho_t(y_t|x) d\rho_t(x) \\ &= \frac{1}{T}\sum_{t=1}^{T}\int_{\mathcal{X}} \Big\langle \psi(f_t(x)), \underbrace{\int_{\mathbb{R}} \phi(y_t)\rho_t(y_t|x)}_{\boldsymbol{g}_t^*(x)} \Big\rangle d\rho_t(x) \end{aligned}$$

The minimizer of the expected risk is then:

$$f^*(x) = \operatorname*{argmin}_{c \in \mathcal{C}} \sum_{t=1}^{T} \langle \psi(c_t), g_t^*(x) \rangle$$

Nonlinear MTL Estimator (I)

Idea: Estimate g_t^* with $\hat{g}_t$ for each $t = 1, ..., T$. Then estimate

$$f^*(x) = \operatorname*{argmin}_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), g_t^*(x) \rangle$$

with

$$\hat{f}(x) = \operatorname*{argmin}_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), \hat{g}_t(x) \rangle$$

Nonlinear MTL Estimator (II)

Let $k : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ be a psd kernel (e.g. the Gaussian kernel). We learn $\hat{g}_t$ via kernel ridge regression:

$$\hat{g}_t = \operatorname*{argmin}_{g \in \mathcal{H}_k} \; \frac{1}{n} \sum_{i=1}^{n} \|g(x_{ti}) - \phi(y_{ti})\|_{\mathcal{H}}^2 + \lambda \|g\|_k^2$$

Then [Thm. 4.1, Micchelli and P., 2005]:

$$\hat{g}_t(x) = \sum_{i=1}^{n} \alpha_{ti}(x) \phi(y_{ti}) \qquad \big(\alpha_{t1}(x), \dots, \alpha_{tn}(x)\big) = (K_t + n\lambda I)^{-1} v_t(x)$$

where $K_t = \big(k(x_{ti}, x_{tj})\big)_{i,j=1}^{n}$ and $v_t(x) = \big(k(x_{ti}, x)\big)_{i=1}^{n}$.

Nonlinear MTL Estimator (III)

Using again the property of the loss

$$\hat{f}(x) = \operatorname*{argmin}_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \langle \psi(c_t), \hat{g}_t(x) \rangle$$

$$= \operatorname*{argmin}_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \left\langle \psi(c_t), \sum_{i=1}^{n} \alpha_{ti}(x) \phi(y_{ti}) \right\rangle$$

$$= \operatorname*{argmin}_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \alpha_{ti}(x) \langle \psi(c_t), \phi(y_{ti}) \rangle$$

$$= \operatorname*{argmin}_{c \in \mathcal{C}} \frac{1}{T} \sum_{t=1}^{T} \sum_{i=1}^{n} \alpha_{ti}(x) \ell(c_t, y_{ti})$$

Note that evaluating $\hat{f}(x)$ does not require knowledge of $\mathcal{H}$, ψ or ϕ!

Nonlinear MTL with Square Loss

- If ℓ is the square loss then

$$\widehat{f}(x) = \underset{c \in \mathcal{C}}{\operatorname{argmin}} \sum_{t=1}^{T} a_t(x)\big(c_t - b_t(x)/a_t(x)\big)^2$$

with

$$a_t(x) = \sum_{i=1}^{n_t} \alpha_{ti}(x), \qquad b_t(x) = \sum_{i=1}^{n_t} \alpha_{ti}(x) y_{ti}$$

- Interpretation: we perform the projection of $\big(b_t(x)/a_t(x)\big)_{t=1}^{T}$ according to the metric induced by the matrix $\operatorname{diag}(a_t(x), ..., a_T(x))$

- If $a_t(x)$ is small it will affect less the weighted projection

Statistical Analysis

Thm. 1 (Comparison inequality).

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq \underbrace{2 \sup_{c \in \mathcal{C}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \|\psi(c_t)\|^2}}_{q_{\mathcal{C},\ell,T}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \|\hat{g}_t - g_t^*\|^2_{\mathcal{L}_2(\rho_X,\mathcal{H})}}$$

Proof idea: Let

$$\bar{\mathcal{E}}(f) = \frac{1}{T} \sum_{t=1}^{T} \int_{\mathcal{X}} \langle \psi(f_t(x)), \hat{g}_t(x) \rangle d\rho_t(x)$$

Then

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) = \underbrace{\mathcal{E}(\hat{f}) - \bar{\mathcal{E}}(\hat{f})}_{A} + \underbrace{\bar{\mathcal{E}}(\hat{f}) - \mathcal{E}(f^*)}_{B}$$

and we can bound A and B with Cauchy Schwarz's inequality

Statistical Analysis (cont.)

Thm. 1 (Comparison inequality).

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq 2 \underbrace{\sup_{c \in \mathcal{C}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \|\psi(c_t)\|^2}}_{q_{\mathcal{C},\ell,T}} \sqrt{\frac{1}{T} \sum_{t=1}^{T} \|\hat{g}_t - g_t^*\|^2_{\mathcal{L}_2(\rho_X, \mathcal{H})}}$$

Implications:

- **Thm. 2 (Consistency).** $\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \to 0 \qquad \text{a.s.}$
- **Thm. 3 (Rates).** If $g_t^* \in \mathcal{H}_k$ for all $t = 1, \dots, T$ then

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \lesssim q_{\mathcal{C},\ell,T} \frac{\log T}{n^{\frac{1}{4}}} \qquad \text{w.h.p}$$

Example

Choose $\ell(y, y') = (y - y')^2$. Then

- If $\mathcal{C}$ is the $T-1$ dimensional sphere then

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq O\big((nT)^{-\frac{1}{4}}\big) \qquad \text{w.h.p.}$$

- In comparison if $\mathcal{C} = [-B, B]^T$ then

$$\mathcal{E}(\hat{f}) - \mathcal{E}(f^*) \leq O\big(n^{-\frac{1}{4}}\big) \qquad \text{w.h.p.}$$

Proof sketch. WLOG we can use the modified loss $\ell(y, y') = y^2 - 2yy'$. Then $\ell(y, z) = \langle \psi(y), \phi(y') \rangle = \langle (y^2, y), (1, -2y') \rangle$. Hence

$$q_{\mathcal{C},\ell,T} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} \|\psi(c_t)\|^2} = \sqrt{\frac{1}{T}\sum_{t=1}^{T} c_t^4 + c_t^2} = \begin{cases} \sqrt{2} & \text{if } \mathcal{C} = [-B, B]^T \\ B\sqrt{\frac{1+B^2}{T}} & \text{if } \mathcal{C} = \{\|c\|_2 \leq B\} \end{cases}$$

Extension: Violating $\mathcal{C}$

- In practice, knowledge of the constraint set $\mathcal{C}$ may not be exact
- One ways to overcome this is to penalize predictions depending on their distance from the set $\mathcal{C}$

$$\mathcal{C}_\delta = \left\{ c + r \ : \ c \in \mathcal{C},\ r \in \mathbb{R}^T,\ \|r\| \leq \delta \right\}$$

  where δ ranges from 0 ($\mathcal{C}_0 = \mathcal{C}$) to $+\infty$ ($\mathcal{C}_\infty = \mathbb{R}^T$).

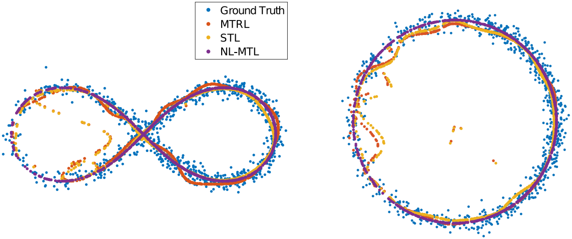
- We can show that

$$\widehat{f_\delta}(x) = \hat{f}(x) + r(x)\ \min(1, \delta / \|r(x)\|)$$

  where $\hat{f}$ is the unperturbed solution and $r = \frac{b(x)}{a(x)} - \hat{f}(x)$

Empirical Results

Synthetic data



Lemniscate $(y_1^4 - (y_1^2 - y_2^2) = 0)$ Circumference

Inverse dynamics (Sarcos)

	STL	MTL[36]	CMTL[10]	MTRL[11]	MTFL[13]	FMTL[16]	NL-MTL[R]	NL-MTL[P]
Expl. Var. (%)	40.5 ±7.6	34.5 ±10.2	33.0 ±13.4	41.6 ±7.1	49.9 ±6.3	50.3 ±5.8	**55.4 ±6.5**	54.6 ±5.1

Ranking (Movielens100k)

	NL-MTL	SELF[21]	Linear [37]	Hinge [38]	Logistic [39]	SVMStruct [20]	STL	MTRL[11]
Rank Loss	**0.271 ±0.004**	0.396 ±0.003	0.430 ±0.004	0.432 ±0.008	0.432 ±0.012	0.451 ±0.008	0.581 0.003	0.613 ±0.005

Open Problems

- Can we improve the error bounds by optimizing over the choice the estimator $\hat{g}$?
- Add further constraints on the problem (e.g. low rankness)
- What if $\mathcal{C}$ is not known a-priori? Can we estimate it?

References

Carlo Ciliberto, Alessandro Rudi, Lorenzo Rosasco, Massimiliano Pontil. Consistent multitask learning with nonlinear output constraints. Advances in Neural Information Processing Systems 30 (NIPS), 2017.

Carlo Ciliberto, Alessandro Rudi, Lorenzo Rosasco. A consistent regularization approach for structured prediction. Advances in Neural Information Processing Systems 29 (NIPS), pages 4412-4420, 2016.

Thomas Hofmann Bernhard Schlkopf Alexander J. Smola Ben Taskar Bakir, Gökhan and S.V.N Vishwanathan. Predicting structured data. MIT press, 2007.

John C Duchi, Lester W Mackey, and Michael I Jordan. On the consistency of ranking algorithms. In Proceedings of the 27th International Conference on Machine Learning (ICML-10), pages 327?334, 2010.

T H A N K   Y O U!

and

H A P P Y   B I R T H D A Y!