Multi-task Learning

Massimiliano Pontil

Department of Computer Science
Centre for Computational Statistics and Machine Learning
University College London

Joint work with Andreas Maurer and Bernardino Romera Paredes
Outline

- Problem formulation and examples
- Sparse coding
- Statistical analysis
- Multilinear multitask learning
- Low rank tensor completion
Problem Formulation

- Fix probability distributions $\mu_1, \ldots, \mu_T$ on $\mathbb{R}^d \times \mathbb{R}$

- Draw data: $z_t = ((x^1_t, y^1_t), \ldots, (x^m_t, y^m_t)) \sim \mu^m_t, \ t = 1, \ldots, T$

- Learn linear predictors $w_1, \ldots, w_T$ by solving

$$\min_{[w_1, \ldots, w_T] \in S} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{m} \sum_{i=1}^{m} \ell(y^i_t, \langle w_t, x^i_t \rangle)$$

where $S$ encourages “common structure” among the tasks.
Problem Formulation (cont.)

\[
\min_{[w_1, \ldots, w_T] \in S} \frac{1}{T} \sum_{t=1}^{T} \frac{1}{m} \sum_{i=1}^{m} \ell(y_t^i, \langle w_t, x_t^i \rangle)
\]

- Example: \( S = \{ \Omega(w_1, \ldots, w_T) \leq \rho \} \)

- Independent task learning (ITL): \( \Omega(w_1, \ldots, w_T) = \max_t \omega(w_t) \)

- Typical scenario: many tasks but only few examples per task
  
  In this regime ITL does not work! [Maurer & P., ALT 2008]
Applications

- **User modelling:**
  - each task is to predict a user’s ratings to products [Lenk et al. 1996,...]
  - the ways different people make decisions about products are related
  - special case (matrix completion): $x_t^i \in \{e_1, \ldots, e_d\}$

- **Multiple object detection in scenes:**
  - detection of each object corresponds to a binary classification task
  - learning common features enhances performance [Torralba et al. 2004,...]

Many more: affective computing, bioinformatics, neuroimaging, NLP, robotics,...
Examples of Regularizers

- Quadratic, e.g. \( \sum_{t=1}^{T} \| w_t \|_2^2 + \frac{1-c}{c} \sum_{t=1}^{T} \| w_t - \bar{w} \|_2^2, \quad c \in (0, 1] \)

- Common sparsity: \( \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{jt}^2} \)

- Common low dimensional subspace: \( \|[w_1, \ldots, w_T]\|_{tr} \)

- Extend to nonlinear model using RKHS!

Encourage \( w_t \)'s which are **sparse combinations** of some vectors:

\[
w_t = D\gamma_t = \sum_{k=1}^{K} D_k\gamma_{kt} : \|\gamma_t\|_1 \leq \alpha
\]

Set of **dictionaries** \( D_K := \{ D = [D_1, ..., D_K] : \|D_k\|_2 \leq 1, \ \forall k \} \)

Learning method [Maurer et al. 2013]:

\[
\min_{D \in D_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\|\gamma\|_1 \leq \alpha} \frac{1}{m} \sum_{i=1}^{m} \ell(\langle D\gamma, x^i_t \rangle, y^i_t)
\]

For fixed \( D \) this is like Lasso with **feature map** \( \phi(x) = D^T x \)
Connection to Sparse Coding

\[
\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\|\gamma\|_1 \leq \alpha} \frac{1}{m} \sum_{i=1}^{m} \ell(\langle D\gamma, x_t^i \rangle, y_t^i)
\]

Natural extension of sparse coding [Olshausen and Field 1996]:

\[
\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\|\gamma\|_1 \leq \alpha} \|w_t - D\gamma\|_2^2
\]

Obtained for \(m \to \infty\), \(\ell\) the square loss and \(y_t^i = \langle w_t, x_t^i \rangle\), \(x_t^i \sim \mathcal{N}(0, I)\)
Experiments

Randomly choose 20 characters from NIST dataset, learn dictionary $\hat{D}$ from all pairwise binary classification tasks, then use $\hat{D}$ on a new set of 10 characters

Tune parameters $K$ and $\alpha$ on a separate set of 10 characters
Learn a dictionary for image reconstruction from few pixel values (input space is the set of possible pixels indices, output space represents the gray level)

Compare resultant dictionary (top) to that obtained by SC (bottom):
Theorem 1. Let $\hat{S}_p := \frac{1}{T} \sum_{t=1}^{T} \| \hat{\Sigma}_t \|_p, \quad p \geq 1$. With probability $\geq 1 - \delta$

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E} [\ell(\langle \hat{D} \gamma_t, x \rangle, y)] - \min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\| \gamma_t \|_1 \leq \alpha} \mathbb{E} [\ell(\langle D \gamma_t, x \rangle, y)]$$

$$\leq L \alpha \sqrt{\frac{8 \hat{S}_\infty \log(2K)}{m}} + L \alpha \sqrt{\frac{2 \hat{S}_1 (K + 12)}{mT}} + \sqrt{\frac{8 \log \frac{4}{\delta}}{mT}}$$

- **Comparable to Lasso** with best a-priori known dictionary! [Kakade et al. 2012]

- If input distribution is uniform on the unit sphere then $\hat{S}_1 = 1$ and $\hat{S}_\infty \approx \frac{1}{m}$

- $O\left(\sqrt{\frac{\log K}{m}}\right)$ vs. $O\left(\sqrt{\frac{K}{m}}\right)$ for trace norm regularization [Maurer & P., 2013]
[Baxter, 2000]: distributions $\mu_1, \ldots, \mu_T \sim \mathcal{E}$ are randomly chosen

Example: $\mu_t(x, y) = p(x)\delta(\langle w_t, x \rangle - y)$, where $w_t$ is random vector

Risk $\mathcal{R}(D) := \mathbb{E}_{\mu \sim \mathcal{E}} \mathbb{E}_{z \sim \mu^m} \mathbb{E}_{(x,y) \sim \mu} \ell(\langle D\gamma(z|D), x \rangle, y)$

Optimal risk $\mathcal{R}^* := \min_{D \in \mathcal{D} \kappa} \mathbb{E}_{\mu \sim \mathcal{E}} \min_{\|\gamma\|_1 \leq \alpha} \mathbb{E}_{(x,y) \sim \mu} \ell(\langle D\gamma, x \rangle, y)$

Theorem 2. Let $S_\infty(\mathcal{E}) := \mathbb{E}_{\mu \sim \mathcal{E}} \mathbb{E}_{z \sim \mu^m} \|\Sigma(x)\|_\infty$. With probability $\geq 1 - \delta$

$$\mathcal{R}(\hat{D}) - \mathcal{R}^* \leq 4L\alpha \sqrt{\frac{S_\infty(\mathcal{E})(2 + \ln K)}{m}} + L\alpha K \sqrt{\frac{2\pi \hat{S}_1}{T}} + \sqrt{\frac{8 \ln \frac{4}{\delta}}{T}}$$
Comparison to Sparse Coding Bound

- Assume: \( \mu_t(x, y) = p(x)\delta(\langle w_t, x \rangle - y) \), with \( w_t \sim \rho \), a prescribed distribution on the unit ball of a Hilbert space.

- Let \( g(w; D) := \min_{\|\gamma\|_1 \leq \alpha} \|w - D\gamma\|_2^2 \)

- Taking \( m \to \infty \) in Theorem 2, we recover a previous bound for sparse coding [Maurer & P., 2010]

\[
\mathbb{E}_{w \sim \rho}[g(w; \hat{D})] - \min_{D \in \mathcal{D}_K} \mathbb{E}_{w \sim \rho}[g(w; D)] \leq 2\alpha(1 + \alpha)K\sqrt{\frac{2\pi}{T}} + \sqrt{\frac{8\ln\frac{4}{\delta}}{T}}
\]
Tasks are identified by a multi-index

Example: predict action-units’ activation (e.g. cheek raiser) for different people: $t = (t_1, t_2) = (\text{“identity”}, \text{“action-unit”})$

[Lucey et. al 2011]
Learn a tensor $\mathbf{W} \in \mathbb{R}^{T_1 \times T_2 \times d}$ from a set of linear measurements $W_{t_1,t_2,:} \in \mathbb{R}^d$ the $(t_1, t_2)$-th regression task, $t_1 = 1, \ldots, T_1$, $t_2 = 1, \ldots, T_2$.

Goal: control rank of each matricization of $\mathbf{W}$:

$$R(\mathbf{W}) := \frac{1}{3} \sum_{n=1}^{3} \text{rank}(\mathbf{W}(n))$$


$$R(\mathbf{W}) \geq \|\mathbf{W}\|_{\text{tr}} := \frac{1}{3} \sum_{n=1}^{3} \|\sigma(\mathbf{W}(n))\|_1$$
Alternative approach using Tucker decomposition

\[ W_{t_1,t_2,j} = \sum_{s_1=1}^{S_1} \sum_{s_2=1}^{S_2} \sum_{k=1}^{p} G_{s_1,s_2,k} A_{t_1,s_1} B_{t_2,s_2} C_{j,k} \]

\[ S_1 \ll T_1, \ S_2 \ll T_2, \ p \ll d \]
Alternative Convex Relaxation

- $\| \cdot \|_{\text{tr}}$ is the tightest convex relaxation of rank on the spectral unit ball [Fazel, Hindi, Boyd, 2001]

\[ \| W \|_{\text{tr}} \leq \text{rank}(W), \quad \forall W \text{ s.t. } \| W \|_{\infty} \leq 1 \]

- Difficulty with tensor setting: $\| W(n) \|_{\infty}$ varies with $n$!

- Relax on Euclidean ball [Romera-Paredes and P. 2013]

\[ \Omega_\alpha(W) = \frac{1}{N} \sum_{n=1}^{N} \omega^{**}_\alpha(\sigma(W(n))) \]

\[ \omega^{**}_\alpha : \text{convex envelop of card}(\cdot) \text{ on the } \ell_2 \text{ ball or radius } \alpha \]

Related work by [Argyriou, Foygel, Srebro, NIPS 2012]
\[ \Omega_\alpha(\mathcal{W}) = \frac{1}{N} \sum_{n=1}^{N} \omega^{**}_\alpha(\sigma(W_n)) \]

**Lemma.** If \( \|x\|_2 = \alpha \) then \( \omega^{**}_\alpha(x) = \text{card}(x) \).

Implication: if \( \exists \mathcal{W} \) s.t. conditions below holds then \( \Omega_{p_{\text{min}}}(\mathcal{W}) > \|\mathcal{W}\|_{\text{tr}} \)

(a) \( \|W(n)\|_\infty \leq 1 \quad \forall n \)

(b) \( \|\mathcal{W}\|_2 = \sqrt{p_{\text{min}}} \)

(c) \( \min_n \text{rank}(W(n)) < \max_n \text{rank}(W(n)) \)

On the other hand, \( \omega^{**}_1 \) is the convex envelope of \( \text{card} \) on \( \ell_2 \) unit ball, so:

\[ \Omega_1(\mathcal{W}) \geq \|\mathcal{W}\|_{\text{tr}}, \quad \forall \mathcal{W} : \|\mathcal{W}\|_2 \leq 1 \]
Problem Reformulation

Want to minimize

$$\frac{1}{\gamma} E(\mathcal{W}) + \sum_{n=1}^{N} \Psi (\mathcal{W}(n))$$

Decouple the regularization term [Gandy et al, 2011; Signoretto et al. 2011]

$$\min_{\mathcal{W}, \mathcal{B}_1, \ldots, \mathcal{B}_N} \left\{ \frac{1}{\gamma} E (\mathcal{W}) + \sum_{n=1}^{N} \Psi (B_{n(n)}) : \mathcal{B}_n = \mathcal{W}, n = 1, \ldots, N \right\}$$

Augmented Lagrangian:

$$\mathcal{L} (\mathcal{W}, \mathcal{B}, \mathcal{C}) = \frac{1}{\gamma} E (\mathcal{W}) + \sum_{n=1}^{N} \left[ \Psi (B_{n(n)}) - \langle \mathcal{C}_n, \mathcal{W} - \mathcal{B}_n \rangle + \frac{\beta}{2} \| \mathcal{W} - \mathcal{B}_n \|^2 \right]$$
\[ \mathcal{L}(W, B, C) = \frac{1}{\gamma} E(W) + \sum_{n=1}^{N} \left[ \psi(B_{n(n)}) - \langle C_n, W - B_n \rangle + \frac{\beta}{2} \| W - B_n \|^2 \right] \]

Updating equations:

\[ W^{[i+1]} \leftarrow \arg\min_W \mathcal{L} \left( W, B^{[i]}, C^{[i]} \right) \]
\[ B_n^{[i+1]} \leftarrow \arg\min_{B_n} \mathcal{L} \left( W^{[i+1]}, B, C^{[i]} \right) \]
\[ C_n^{[i+1]} \leftarrow C_n^{[i]} - \left( \beta W^{[i+1]} - B_n^{[i+1]} \right) \]

- 2nd step involves the computation of proximity operator of \( \Psi \)
Let $B = B_{n(n)}$ and where $A = (\mathcal{W} - \frac{1}{\beta} \mathcal{C}_n)(n)$. Rewrite 2nd step as:

$$\hat{B} = \text{prox}_{\frac{1}{\beta} \psi}(A) := \arg\min_B \left\{ \frac{1}{2} \|B - A\|_2^2 + \frac{1}{\beta} \Psi(B) \right\}$$

Case of interest: $\Psi(B) = \psi(\sigma(B))$

By von Neuman’s inequality:

$$\text{prox}_{\frac{1}{\beta} \psi}(A) = U_A \text{diag} \left( \text{prox}_{\frac{1}{\beta} \psi}(\sigma_A) \right) V_A^\top$$

If $\psi(x) = \omega^{**}_{\alpha}$ use $\text{prox}_{\frac{1}{\beta} \omega^{**}_{\alpha}}(x) = x - \frac{1}{\beta} \text{prox}_{\beta \omega^{*}_{\alpha}}(\beta x)$

$$\omega^{*}_{\alpha}(z) = \sup_{\|x\|_2 \leq \alpha} \{ \langle x, z \rangle - \text{card}(x) \} = \max_{0 \leq r \leq d} (\alpha \|z_{1:r}\|_2 - r)$$
Experiments

Video compression (Left) and exam score prediction (Right):

Time comparison:
MTL exploits relationships between multiple learning tasks to improve over independent task learning under specific conditions.

Method to learn a dictionary for sparse coding of multiple tasks. Matches performance of Lasso with a-priori known dictionary.

Multilinear MTL: need for convex regularizers which encourage low rank tensors.
References

[Cavallanti, Cesa-Bianchi, Gentile] Linear algorithms for online multitask classification, JMLR 2010.
[Kakade, Shalev-Shwartz, Tewari] Regularization techniques for learning with matrices, JMLR 2012.
[Kang, Grauman, Sha] Learning with Whom to Share in Multi-task Feature Learning. ICML 2011.
[Kumar and Daumé III] Learning Task Grouping and Overlap in Multi-task LearningAbstract, ICML 2012