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: $\mathbf{z}_t = \big((x_t^1, y_t^1), \ldots, (x_t^m, y_t^m)\big) \sim \mu_t^m, \quad t = 1,\ldots,T$

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

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

- Set $\mathcal{S}$ encourages "common structure" among the tasks

Problem Formulation (cont.)

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

- Example: $\mathcal{S} = \{\Omega(w_1, \ldots, w_T) \leq \rho\}$
- 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. $\sum_{t=1}^{T} \|w_t\|_2^2 + \frac{1-c}{c} \sum_{t=1}^{T} \left\|w_t - \bar{w}\right\|_2^2$, $c \in (0,1]$
- Common sparsity: $\sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{jt}^2}$
- Common low dimensional subspace: $\|[w_1, \ldots, w_T]\|_{\mathrm{tr}}$
- Extend to nonlinear model using RKHS!

[Argyriou et al. 2006, 2008, 2009; Baldassarre et al. 2012; Caponnetto et al. 2008; Carmeli et al. 2006; Cavallanti et al. 2009; Dinuzzo & Fukumizu, 2012; Evgeniou & P. 2004; Evgeniou et al. 2005; Jacob et al. 2008; Koltchinskii et al. 2011; Kumar & Daumé III, 2012; Lounici et al., 2009, 2011; Maurer, 2006; Micchelli & P., 2005; Obozinski et al. 2009; Romera-Paredes et al. 2012; Salakhutdinov et al, 2011,...]

Learning Sparse Representations

- Encourage w_t's which are **sparse combinations** of some vectors:

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

- Set of **dictionaries** $\mathcal{D}_K := \left\{ D = [D_1, ..., D_K] \ : \ \|D_k\|_2 \leq 1, \ \forall k \right\}$

- Learning method [Maurer et al. 2013]:

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

- For fixed D this is like Lasso with **feature map** $\phi(x) = D^\top x$

Connection to Sparse Coding

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

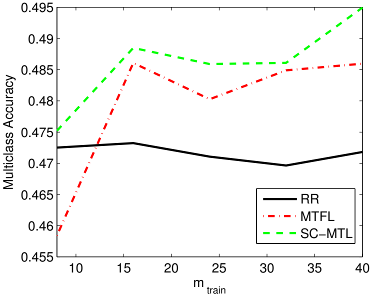
Natural extension of sparse coding [Olshausen and Field 1996]:

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

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

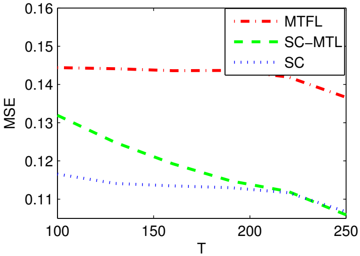
Experiments

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

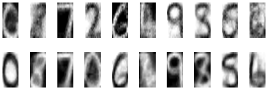


Tune parameters K and α on a separate set of 10 characters

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 $\widehat{S}_p := \frac{1}{T}\sum_{t=1}^{T} \|\widehat{\Sigma}_t\|_p$, $p \geq 1$. With probability $\geq 1-\delta$

$$\frac{1}{T}\sum_{t=1}^{T} \mathop{\mathbb{E}}_{(x,y)\sim\mu_t} \ell(\langle \widehat{D}\hat{\gamma}_t, x\rangle, y) - \min_{D\in\mathcal{D}_K} \frac{1}{T}\sum_{t=1}^{T} \min_{\|\gamma_t\|_1\leq\alpha} \mathop{\mathbb{E}}_{(x,y)\sim\mu_t} \ell(\langle D\gamma_t, x\rangle, y)$$

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

- **Comparable to Lasso** with best a-priori known dictionary! [Kakade et al. 2012]
- If input distribution is uniform on the unit sphere then $\widehat{S}_1 = 1$ and $\widehat{S}_\infty \approx \frac{1}{m}$
- $O\left(\sqrt{\frac{\log K}{m}}\right)$ 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 $\mathcal{R}(D) := \underset{\mu\sim\mathcal{E}}{\mathbb{E}} \ \underset{\mathbf{z}\sim\mu^m}{\mathbb{E}} \ \underset{(x,y)\sim\mu}{\mathbb{E}} \ell(\langle D\gamma(\mathbf{z}|D), x\rangle, y)$

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

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

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

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

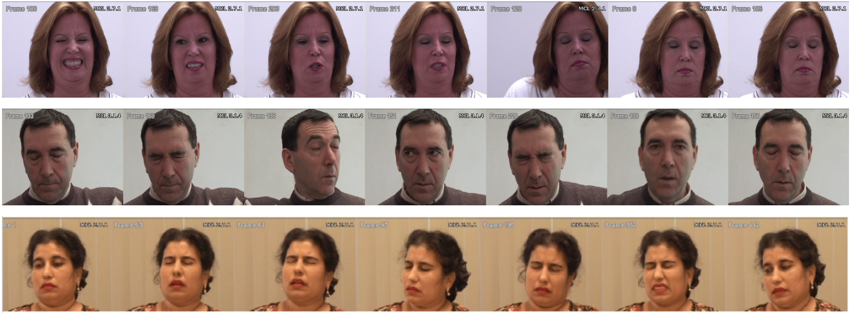
- Taking $m \to \infty$ in Theorem 2, we recover a previous bound for sparse coding [Maurer & P., 2010]

$$\mathop{\mathbb{E}}_{w\sim\rho}\left[g(w; \widehat{D})\right] - \min_{D\in\mathcal{D}_K} \mathop{\mathbb{E}}_{w\sim\rho}\left[g(w; D)\right] \leq 2\alpha(1+\alpha)K\sqrt{\frac{2\pi}{T}} + \sqrt{\frac{8\ln\frac{4}{\delta}}{T}}$$

Multilinear MTL

[Romera-Paredes et al. 2013]

- Tasks are identified by a multi-index
- Example: predict action-units' activation (e.g. `cheek raiser`) for different people: $t = (t_1, t_2) =$ ("identity", "action-unit")



[Lucey et. al 2011]

Multilinear MTL (cont.)

- Learn a tensor $\mathcal{W} \in \mathbb{R}^{T_1 \times T_2 \times d}$ 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 W:

$$R(\mathcal{W}) := \frac{1}{3} \sum_{n=1}^{3} \mathrm{rank}(W_{(n)})$$

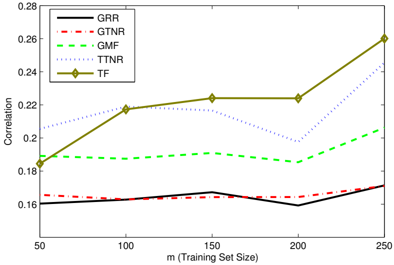
- Convex relaxation [Liu et al. 2011, Gandy et al. 2011, Signoretto et al. 2012]

$$R(\mathcal{W}) \geq \|\mathcal{W}\|_{\mathrm{tr}} := \frac{1}{3} \sum_{n=1}^{3} \|\sigma(W_{(n)})\|_1$$

- 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

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

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

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

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

$$\Omega_\alpha(\mathcal{W}) = \frac{1}{N}\sum_{n=1}^{N} \omega_\alpha^{**}(\sigma(W_{(n)}))$$

ω_α^{**} : convex envelop of $\mathrm{card}(\cdot)$ on the ℓ_2 ball or radius α

Related work by [Argyriou, Foygel, Srebro, NIPS 2012]

Quality of Relaxation (cont.)

$$\Omega_\alpha(\mathcal{W}) = \frac{1}{N}\sum_{n=1}^{N} \omega_\alpha^{**}(\sigma(W_{(n)}))$$

Lemma. If $\|x\|_2 = \alpha$ then $\omega_\alpha^{**}(x) = \mathrm{card}(x)$.

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

(a) $\|W_{(n)}\|_\infty \leq 1\ \forall n$

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

(c) $\min_n \mathrm{rank}(W_{(n)}) < \max_n \mathrm{rank}(W_{(n)})$

On the other hand, ω_1^{**} is the convex envelope of card on ℓ_2 unit ball, so:

$$\Omega_1(\mathcal{W}) \geq \|\mathcal{W}\|_{\mathrm{tr}}, \quad \forall\, \mathcal{W} : \|\mathcal{W}\|_2 \leq 1$$

Problem Reformulation

Want to minimize

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

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

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

Augmented Lagrangian:

$$\mathcal{L}\left(\mathcal{W},\mathcal{B},\mathcal{C}\right) = \frac{1}{\gamma}E\left(\mathcal{W}\right) + \sum_{n=1}^{N}\left[\Psi\left(B_{n(n)}\right) - \langle \mathcal{C}_n, \mathcal{W} - \mathcal{B}_n\rangle + \frac{\beta}{2}\left\|\mathcal{W} - \mathcal{B}_n\right\|_2^2\right]$$

ADMM

$$\mathcal{L}(\mathcal{W},\mathcal{B},\mathcal{C}) = \frac{1}{\gamma}E(\mathcal{W}) + \sum_{n=1}^{N}\left[\Psi\Big((B_{n(n)})\Big) - \langle \mathcal{C}_n, \mathcal{W} - \mathcal{B}_n\rangle + \frac{\beta}{2}\|\mathcal{W} - \mathcal{B}_n\|_2^2\right]$$

Updating equations:

$$\begin{aligned}
\mathcal{W}^{[i+1]} &\leftarrow \operatorname*{argmin}_{\mathcal{W}} \mathcal{L}\left(\mathcal{W}, \mathcal{B}^{[i]}, \mathcal{C}^{[i]}\right) \\
\mathcal{B}_n^{[i+1]} &\leftarrow \operatorname*{argmin}_{\mathcal{B}_n} \mathcal{L}\left(\mathcal{W}^{[i+1]}, \mathcal{B}, \mathcal{C}^{[i]}\right) \\
\mathcal{C}_n^{[i+1]} &\leftarrow \mathcal{C}_n^{[i]} - \left(\beta\mathcal{W}^{[i+1]} - \mathcal{B}_n^{[i+1]}\right)
\end{aligned}$$

- 2nd step involves the computation of proximity operator of Ψ

Proximity Operator

Let $B = B_{n(n)}$ and where $A = (\mathcal{W} - \frac{1}{\beta}\mathcal{C}_n)_{(n)}$. Rewrite 2nd step as:

$$\hat{B} = \mathrm{prox}_{\frac{1}{\beta}\Psi}(A) := \operatorname*{argmin}_{B} \left\{ \frac{1}{2} \|B - A\|_2^2 + \frac{1}{\beta}\Psi(B) \right\}$$

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

By von Neuman's inequality:

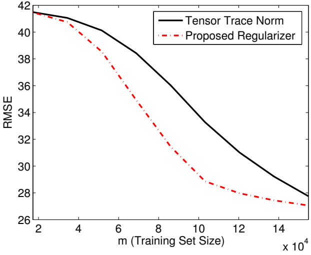
$$\mathrm{prox}_{\frac{1}{\beta}\Psi}(A) = U_A \mathrm{diag}\Big(\mathrm{prox}_{\frac{1}{\beta}\psi}(\sigma_{\mathrm{A}})\Big) V_A^\top$$

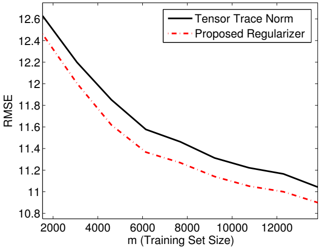
If $\psi(x) = \omega_\alpha^{**}$ use $\mathrm{prox}_{\frac{1}{\beta}\omega_\alpha^{**}}(x) = x - \frac{1}{\beta}\mathrm{prox}_{\beta\omega_\alpha^*}(\beta x)$

$$\omega_\alpha^*(z) = \sup_{\|x\|_2 \le \alpha} \{\langle x, z\rangle - \mathrm{card}(x)\} = \max_{0 \le r \le d} (\alpha \|z_{1:r}^{\downarrow}\|_2 - r)$$

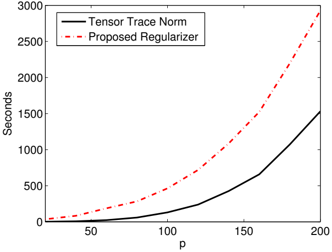
Experiments

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





Time comparison:



Conclusions

- 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

[Argyriou, Evgeniou, Pontil] **Convex multi-task feature learning.** Machine Learning 2008.

[Argyriou, Maurer, Pontil] **An algorithm for transfer learning in a heterogeneous environment**. ECML 2008b.

[Baxter] **A model for inductive bias learning.** JAIR 2000.

[Caponnetto, Micchelli, Pontil, Ying] **Universal mult-task kernels.** JMLR 2008.

[Caruana] **Multi–task learning.** Machine Learning 1998.

[Cavallanti, Cesa-Bianchi, Gentile] **Linear algorithms for online multitask classification**, JMLR 2010.

[Dinuzzo and Fukumizu] **Learning low-rank output kernels.** ACML 2011.

[Evgeniou and Pontil] **Regularized multi-task learning.** SIGKDD 2004.

[Evgeniou, Micchelli, Pontil] **Learning multiple tasks with kernel methods.** JMLR 2005.

[Fazel, Hindi and Boyd] **A rank minimization heuristic with application to minimum order system approximation.** American Control Conference, 2001.

[Gandy, Recht, Yamada] **Tensor completion and low-n-rank tensor recovery via convex optimization.** Inverse Problems, 2011.

[Jacob, Bach, Vert] **Clustered multi-task learning: a convex formulation.** NIPS 2008.

[Lounici, Pontil, Tsybakov, van de Geer] **Oracle inequalities and optimal inference under group sparsity.** Annals of Stat., 2011.

[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

[Lenk, DeSarbo, Green, Young] **Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs.** Marketing Science 1996.

[Liu, Musialski, Wonka, Ye] **Tensor completion for estimating missing values in visual data.** ICCV 2009.

[Maurer] **Bounds for linear multi-task learning.** JMLR 2006.

[Maurer and Pontil] **K-dimensional coding schemes in Hilbert spaces** IEEE Transactions on Information Theory, 2010.

[Maurer, Pontil, Romera-Paredes] **Sparse coding for multitask and transfer learning.** ICML 2013.

[Micchelli and Pontil] **On learning vector-valued functions.** Neural Computation 2005.

[Obozinski, Taskar, Jordan] **Joint covariate selection and joint subspace selection for multiple classification problems.** Statistics and Computing, 2010.

[Romera-Paredes, Aung, Bianchi-Berthouze, Pontil] **Multilinear mulittask learning.** ICML 2013.

[Romera-Paredes and Pontil] **A new convex relaxation for tensor completion.** Preprint, 2013.

[Signoretto, Van de Plas, De Moor, Suykens] **Tensor versus matrix completion: a comparison with application to spectral data.** IEEE Signal Processing Letters, 2011.