Multi-task Learning and Structured Sparsity

Massimiliano Pontil

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

Outline

- Problem formulation and examples
- Design of regularization functions
- Statistical analysis
- Optimization methods
- Extensions

Problem formulation

- Let $\mu_1, \ldots, \mu_T$ be prescribed probability measures on $X \times Y$
- $(x_{t1}, y_{t1}), \ldots, (x_{tn}, y_{tn}) \sim \mu_t$, $t = 1, \ldots, T$
- Goal: find functions $f_t : X \to Y$ which minimize

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim\mu_t} L(y, f_t(x))^2$$

- Regularization approach:

$$\min \ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} L(y_{ti}, f_t(x_{ti})) + \lambda \ \Omega(f_1, \ldots, f_T)$$

- Penalty Ω encourages "common structure" among the functions

Problem formulation (cont.)

- Focus on linear regression

$$\min \frac{1}{T}\sum_{t=1}^{T}\underbrace{\frac{1}{n}\sum_{i=1}^{n} L(y_{ti}, w_t{}^{\top} x_{ti})}_{\text{training error task t}} + \lambda \underbrace{\Omega(w_1, \ldots, w_T)}_{\text{joint regularizer}}$$

- Single task learning: $\Omega(w_1, \ldots, w_T) = \sum_t \omega(w_t)$
- Typical scenario: **many tasks** but only **few examples per task**
- If the tasks are "related", learning them *jointly* should improve over learning each task *independently*

Example 1: user modeling

- Each task is to predict a user's ratings to products

CPU	CD	RAM	$\cdots$	HD	Screen	Price	Rating
1GHz	Y	1GB	$\cdots$	40G	15in	\$1000	7
1GHz	N	1.5GB	$\cdots$	20G	13in	\$1200	3
1.5GHz	Y	1.5GB	$\cdots$	40G	17in	\$1700	5
2GHz	Y	2GB	$\cdots$	80G	15in	\$2000	?
1.5GHz	N	2GB	$\cdots$	40G	13in	\$1800	?

- The ways different people make decisions about products are related, e.g. small variance of parameters

Example 2: matrix completion / collaborative filtering

- Estimate the entries of a products/users matrix (e.g. Netflix)
  $W := [w_1, \ldots, w_T]$

5	?	?	5	?
?	2	3	?	5
?	1	?	?	3
4	?	5	?	?
?	?	1	2	?

- Special case of MTL with $X = \{e_1, \ldots, e_d\}$, $x_{ti} = e_{k(ti)}$:

$$\min_W \frac{1}{Tn} \sum_{t=1}^{T} \sum_{i=1}^{n} (y_{ti} - W_{t,k(ti)})^2 + \lambda \, \Omega(W)$$

Example 3: object detection

- Multiple object detection in scenes: detection of each object corresponds to a binary classification task

- Learning common visual features enhances performance Early work in

  ML used a hidden layer neural nets with hidden weights shared by all the tasks [Baxter 96, Caruana 97, Silver and Mercer 96, etc.]

Quadratic regularizer

[Evgeniou et al. 2005, Caponnetto et al. 08, Baldassarre et al. 10,...]

$$\min \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} L(y_{ti}, w_t{}^\top x_{ti}) + \lambda \, \Omega(w_1, \ldots, w_T)$$

- Let $\Omega(w) = w^\top E w$, with $E \in \mathbf{S}_{++}^{dT \times dT}$, $w = (w_1, \ldots, w_T) \in \mathbb{R}^{dT}$
- Encourages closeness of task parameters / linear relationships – if block diagonal, tasks are learned *independently*
- Example $\Omega(w) = \sum_{t=1}^{T} \|w_t\|^2 + \frac{1-\gamma}{\gamma} \sum_{t=1}^{T} \|w_t - \frac{1}{T} \sum_{s=1}^{T} w_s\|^2$

  $\gamma \in [0, 1]$, $\gamma = 1$: independent tasks, $\gamma = 0$: identical tasks

Equivalent formulation

- Consider function $(x, t) \mapsto f_t(x) := v^\top B_t x$, for $v \in \mathbb{R}^p$ and $B_t \in \mathbb{R}^{p \times d}$ (task specific)
- Rewrite optimization problem as:

$$S(v) = \sum_{t=1}^{T} \sum_{i=1}^{n} L(y_{ti}, v^\top B_t x_{ti}) + \lambda v^\top v$$

- Previous example:

$$B_t^\top = [(1-\gamma)^{\frac{1}{2}} \mathbf{I}_{d \times d}, \underbrace{\mathbf{0}_{d \times d}, \ldots, \mathbf{0}_{d \times d}}_{t-1}, (\gamma T)^{\frac{1}{2}} \mathbf{I}_{d \times d}, \underbrace{\mathbf{0}_{d \times d}, \ldots, \mathbf{0}_{d \times d}}_{T-t}]$$

- Interpretation:

$$w_t = B_t{}^\top v = \sqrt{1-\gamma} v_0 + \sqrt{\gamma T} v_t = \text{“}\textit{common}\text{”} + \text{“}\textit{task specific}\text{”}$$

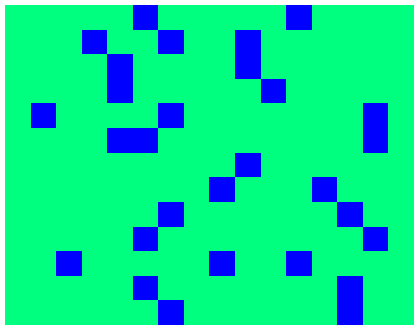
Structured sparsity: few shared variables

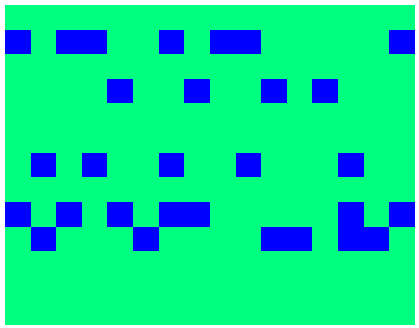
- Favour matrices with many zero rows:

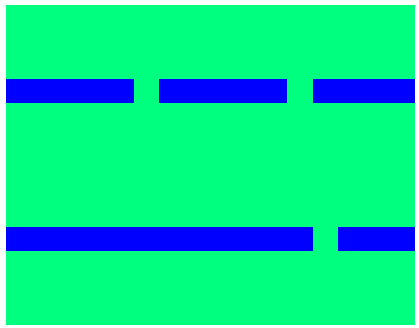
$$\|W\|_{2,1} := \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{jt}^2}$$

- Special case of **group Lasso** method [Lounici et al. 09, Obozinski et al. 10, Yuan and Lin, 06]

Compare matrices W favoured by different regularizers ($\texttt{green} = 0$, $\texttt{blue} = 1$):







#rows =	13	5	2
$\|\cdot\|_{2,1}$ =	19	12	8
ℓ_1-norm =	29	29	29

Statistical analysis

- Linear regression model: $y_{ti} = w_t{}^\top x_{ti} + \epsilon_{ti}$, with ϵ_{ti} i.i.d. $N(0, \sigma^2)$ $i = 1, \ldots, n$, $d \gg n$, use the square loss: $L(y, y') = (y - y')^2$.
- Assume $\text{card}\left\{ j \,:\, \sum_{t=1}^{T} w_{jt}^2 > 0 \right\} \leqslant s$
- Variables not too correlated: $\frac{1}{n} \left| \sum_{i=1}^{n} (x_{ti})_j (x_{ti})_k \right| \leq \frac{1-\rho}{7s}, \ \forall t, \ \forall j \neq k$

Theorem [Lounici et al. 2011] If $\lambda = \frac{4\sigma}{\sqrt{nT}} \sqrt{1 + A \frac{\log d}{T}}$, $A \geq 4$ then w.h.p.

$$\frac{1}{T} \sum_{t=1}^{T} \| \hat{w}_t - w_t \|^2 \ \leq \ \left(\frac{c\sigma}{\rho} \right)^2 \frac{s}{n} \sqrt{1 + A \frac{\log d}{T}}$$

- Dependency on the dimension d is *negligible* for large T
- Compare to Lasso: $\frac{1}{T} \sum_{t=1}^{T} \| w_t^{(L)} - w_t \|^2 \geq c' \frac{s}{n} \log(d \ T)$

Multitask feature learning [Argyriou et al., 2007]

Extend above formulation to learn a low dimensional representation:

$$\min_{U,A} \left\{ \sum_{t=1}^{T} \sum_{i=1}^{n} L(y_{ti}, a_t{}^\top U^\top x_{ti}) + \lambda \, \|A\|_{2,1} : U^\top U = I_{d\times d}, \; A \in \mathbb{R}^{d\times T} \right\}$$

- Let $W = UA$ and minimize over orthogonal U

$$\min_{U} \|U^\top W\|_{2,1} = \|W\|_{\mathrm{tr}} := \sum_{j=1}^{r} \sigma_j(W)$$

Obtain trace norm regularization

$$\min_{W} \sum_{t=1}^{T} \sum_{i=1}^{n} L(y_{ti}, w_t{}^\top x_{ti}) + \lambda \, \|W\|_{\mathrm{tr}}$$

Variational form and alternate minimization

- **Lemma**: $\|W\|_{\mathrm{tr}} = \frac{1}{2} \inf_{D \succ 0} \left\{ \mathrm{tr}(D^{-1} W W^\top) + \mathrm{tr}(D) \right\}$ where the infimizer is $D(W) = (W W^\top)^{\frac{1}{2}}$

$$\min_{W,\ D \succ 0} \sum_{t=1}^{T} \sum_{i=1}^{n} L(y_{ti}, w_t{}^\top x_{ti}) + \frac{\lambda}{2} \Big[\underbrace{\mathrm{tr}(W^\top D^{-1} W)}_{\sum_{t=1}^{T} w_t{}^\top D^{-1} w_t} + \mathrm{tr}(D) \Big]$$

- Further constraining D to be diagonal yields $\|W\|_{2,1}$
- Extension (spectral regularizers) e.g. Schatten p-norms
- Requires a perturbation step in order to prove convergence
- See [Dudík et al, 2012] for comparison results

Risk bound

Theorem [Maurer and P. 2012] Let $R(W) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim\mu_t} L(y, w_t{}^\top x)$ and $\hat{R}(W)$ the empirical error. Assume $L(y,\cdot)$ is ϕ-Lipschitz and $\|x_{ti}\| \le 1$. If $\hat{W} \in \operatorname{argmin}\left\{\hat{R}(W) : \|W\|_{tr} \le B\sqrt{T}\right\}$ then with prob. at least $1-\delta$

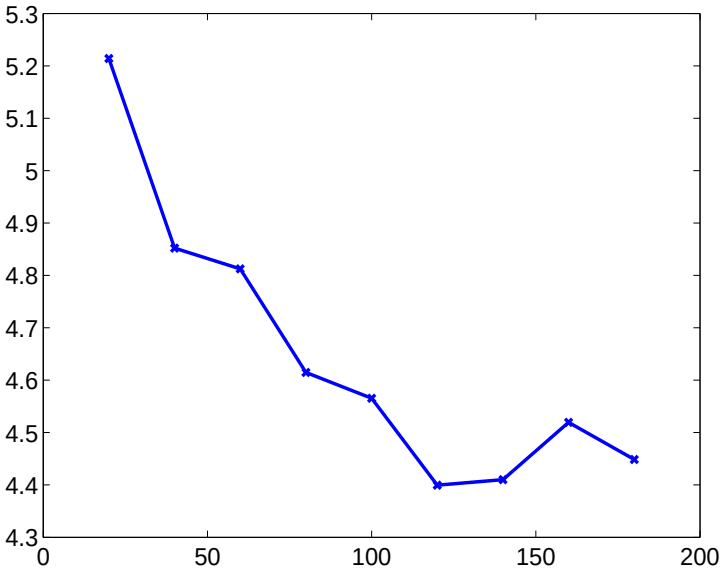
$$R(\hat{W}) - R(W^*) \le 2\phi B\left(\sqrt{\frac{\|\hat{C}\|_\infty}{n}} + \sqrt{\frac{2\left(\ln(nT)+1\right)}{nT}}\right) + \sqrt{\frac{8\ln(3/\delta)}{nT}}$$

with $\hat{C} = \frac{1}{nT}\sum_{t,i} x_{ti} x_{ti}{}^\top$ and $W^* \in \operatorname{argmin} R(W)$

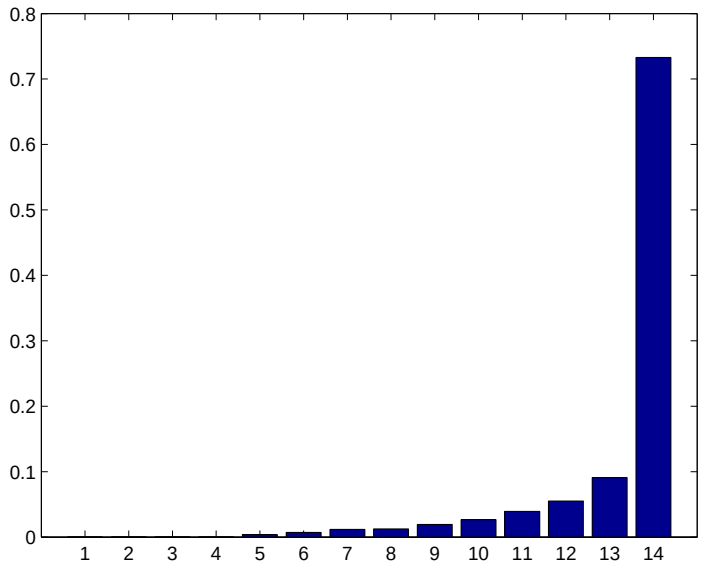
- **Interpretation:** Assume $\operatorname{rank}(W^*) = K$, $\|w_t^*\| \le 1$ and choose $B = K^{1/2}$. If the inputs are uniformly distributed, as T grows we have a $O(\sqrt{K/nd})$ bound as compared to $O(\sqrt{1/n})$ for single task learning

Experiment (computer survey)

Test error vs. #tasks



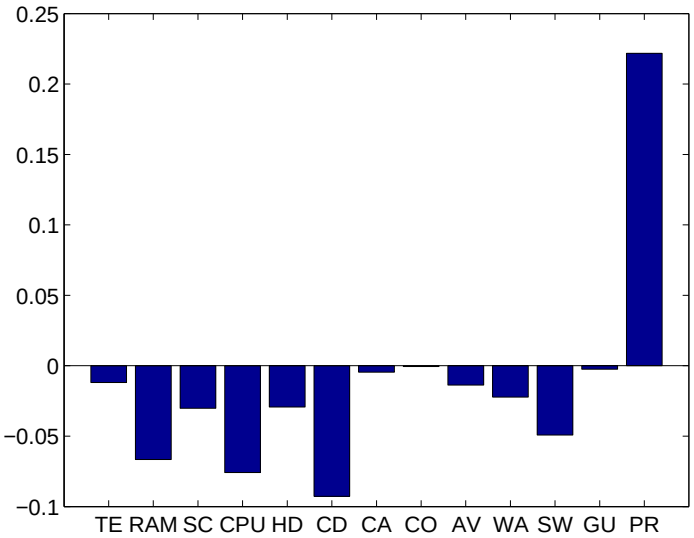
Eigenvalues of matrix D



- Performance improves with more tasks
- A single most important feature shared by everyone

Dataset [Lenk et al. 1996]: consumers' ratings of PC models: 180 persons (tasks), 8 training, 4 test points, 13 inputs (RAM, CPU, price etc.), output in $\{0, \ldots, 10\}$ (likelihood of purchase)

Experiment (computer survey)



Method	Test
Independent	15.05
Aggregate	5.52
Quadratic (best λ)	4.37
Structured Sparsity	4.04
Trace norm	3.72
Quadratic + Trace	3.20

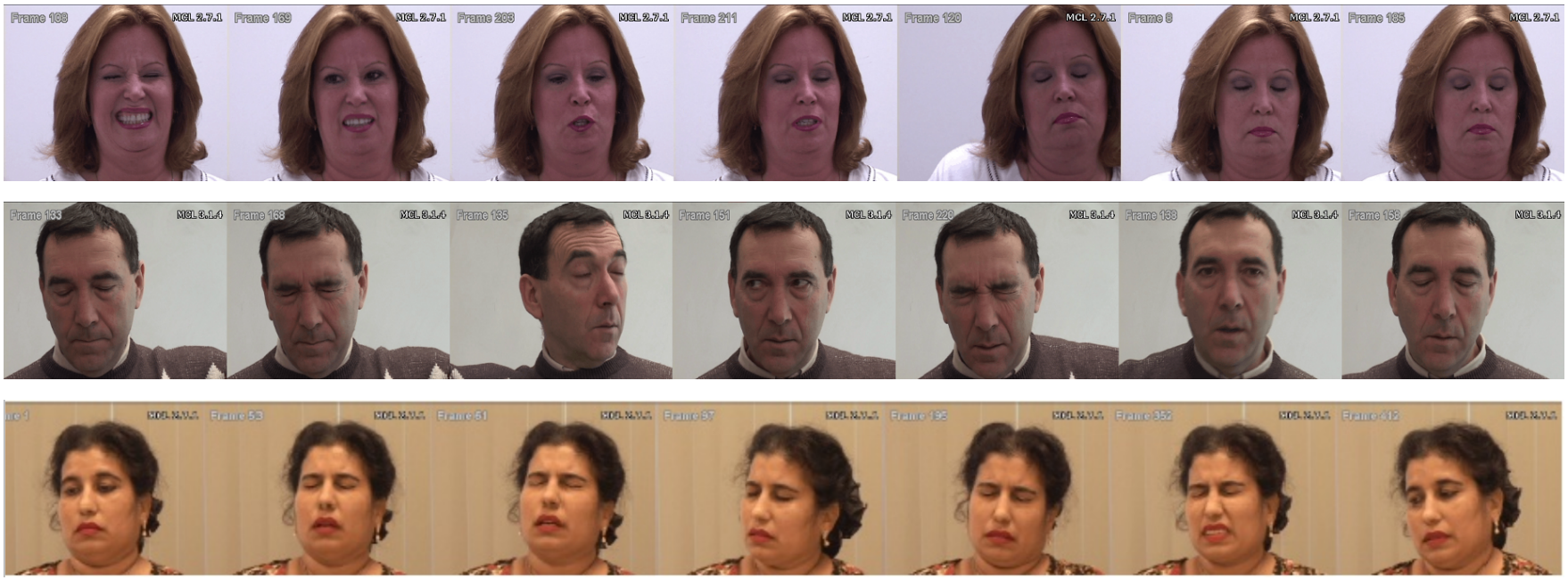
- The most important feature (1st eigenvector of D) weighs *technical characteristics* (RAM, CPU, CD-ROM) vs. *price*

More complex models

- Composite regularizers: $\Omega(B \circ W)$, e.g. $\Omega([w_1 - \bar{w}, \ldots, w_T - \bar{w}])$. More challenging optimization problem [Argyriou et al. 2011]
- "Robust" model: $\Omega(W) = \min_{W=V+Z} \Omega(V) + \|Z^\top\|_{2,1}$ [Chen et al. 2011]
- Constrained variational [Micchelli, Morales, P., 2010], e.g. $\mathcal{D} = \{\mathrm{diag}(\lambda_1, \ldots, \lambda_d) : \lambda \in \Lambda\}$, with $\Lambda \subseteq \mathbb{R}^d_{++}$ a convex cone
- Multitask clustering [Jacob et al. 2008]
- Tensor learning [Gandy et al. 2011, Liu et al. 2011, Signoretto et al. 2012]
- Encourage heterogeneous features [Romera-Paredes et al., 2012]
- Sparse coding for MTL [Kumar and Daumé III, 2012, Maurer et al. 2012]

Tensor learning

Example: predict action units' (e.g. `cheek raiser`) activation for different people [Lucey et. al 2011]



Now t is a double index corresponding to identity/action unit

Tensor learning (cont.)

Let $\mathbf{W} \in \mathbb{R}^{d \times T_1 \times T_2}$. For all $t_1 \in [1{:}T_1]$, $t_2 \in [1{:}T_2]$, $W_{:,t_1,t_2} \in R^d$ is a regression vector from which we generate data samples

- Goal: control rank of each matricization of W:

$$\mathrm{rank}(W_{(1)}) + \mathrm{rank}(W_{(2)}) + \mathrm{rank}(W_{(3)})$$

  where $W_{(n)}$ is the mode-n matricization of $\mathbf{W}$

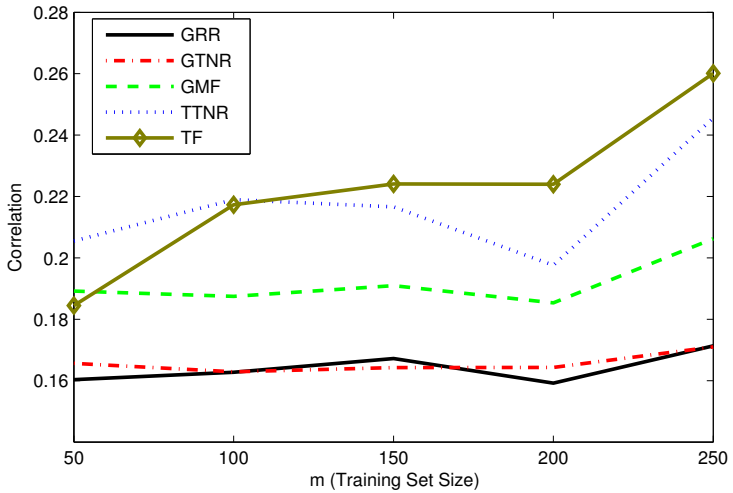
- Convex lower bound [Liu et al. 2011, Gandy et al 2011, Signoretto et al. 2012]

$$\sum_{n=1}^{3} \|W_{(n)}\|_{\mathrm{tr}}$$

Tensor learning (cont.)

$$E(\mathcal{L}(\mathbf{W})) + \lambda \sum_{n=1}^{3} \|W_{(n)}\|_{\mathrm{tr}}$$

- Solved by alternating direction of multipliers [Gandy et al. 2011]
- Alternative non-convex approach using Tucker decomposition

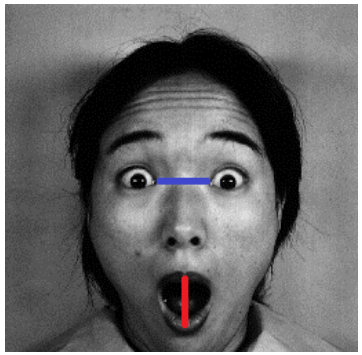


- Convex lower bound not tight – ongoing work studying alternative convex relaxations [Argyriou et al. 2012]

Exploiting unrelated groups of tasks

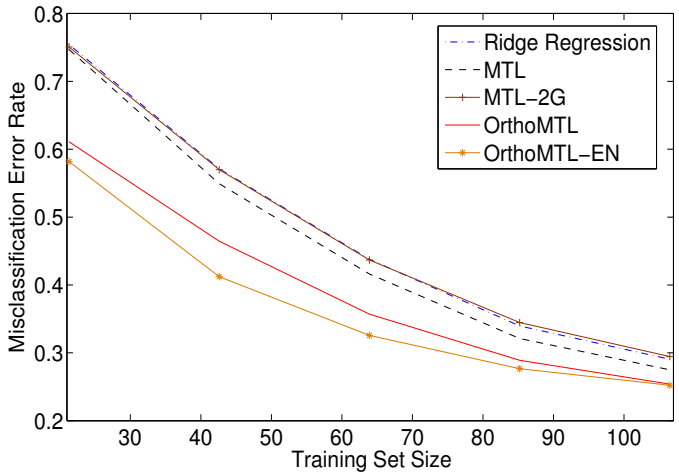
Example: recognizing identity and emotion on a set of faces

- emotion related feature
- identity related feature



Assumption:

1. Low rank within each group
2. Tasks from different groups tend to use orthogonal features



$$\min \left\{ \mathrm{Err}(W) + \mathrm{Err}(V) + \gamma \left[\|[W, V]\|_{\mathrm{tr}} + \rho \|W^\top V\|_{\mathrm{Fr}}^2 \right] \right\}$$

- Related convex problem under conditions [Romera-Paredes et al., 2012]

Multi-task learning with dictionaries

[Maurer et al. 2012]

- Method (natural extension of [Olshausen and Field 1996])

$$\min \left\{ \frac{1}{T} \sum_{t=1}^{T} \|y_t - X_t w_t\|_2^2 : w_t = U a_t \right\}$$

- $U = [u_1, \ldots, u_K]$ with $\|u_k\|_2 \leq 1$
- Similar approach with Frobenius norm bound [Kumar and Daumé III]
- Sparse coding constraint: $\|a_t\|_1 \leq \alpha$ or other structure sparsity norms [Jenatton et al. 2011]
- Estimation bounds show improvement over single task learning and trace norm regularization

Conclusions

- Multi-task learning is ubiquitous – exploiting task relatedness can enhance learning performance
- Reviewed families of regularizers which naturally extend complexity notions (smoothness and sparsity) used for the single-task learning; touched upon statistical analyses and optimisation techniques
- Wide scope for convex relaxation and optimisation techniques in more complex task-relatedness scenarios
- Further work on nonlinear MTL via reproducing kernel Hilbert spaces

Thanks

- Andreas Argyriou
- Nadia Bianchi-Berthouze
- Andrea Caponnetto
- Theodoros Evgeniou
- Karim Lounici
- Andreas Maurer
- Charles Micchelli
- Bernardino Romera-Paredes
- Alexandre Tsybakov
- Sara van de Geer
- Yiming Ying

References (I)

[Abernethy, Bach, Evgeniou, Vert] **A new approach to collaborative filtering: operator estimation with spectral regularization.** JMLR 2009.

[Ando and Zhang] **A framework for learning predictive structures from multiple tasks and unlabeled data.** JMLR 2005.

[Argyriou, Evgeniou, Pontil] **Multi-task feature learning.** NIPS 2006.

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

[Argyriou, Foygel, Srebro] **Sparse Prediction with the k-Support Norm**. NIPS 2012.

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

[Argyriou, Micchelli, Pontil] **When is there a representer theorem? Vector versus matrix regularizers.** JMLR 2009.

[Argyriou, Micchelli, Pontil, Shen, Xu] **Efficient first order methods for linear composite regularizers.** arXiv:1104.1436.

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

[Ben-David and Schuller] **Exploiting task relatedness for multiple task learning.** COLT 2003.

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

[Carmeli, De Vito, Toigo] **Vector valued reproducing kernel Hilbert spaces, integrable functions and Mercer theorem.** Analysis and Applications, 2006.

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

References (II)

[Dudík, Harchaoui, Malik] **Lifted coordinate descent for learning with trace-norm regularization**, AISTATS 2012.

[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.

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

[Jebara] **Multi-task feature and kernel selection for SVMs.** ICML 2004.

[Lounici, Pontil, Tsybakov, van de Geer] **Taking advantage of sparsity in multi-task learning.** COLT 2009.

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

[Izenman] **Reduced-rank regression for the multivariate linear model**, J. Multivariate Analysis, 1975.

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

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

[Maurer and Pontil] **Structured sparsity and generalization.** JMLR 2012.

[Maurer, Pontil, Romera-Paredes] **Sparse coding for multitask and transfer learning.** arXiv:1209.0738.

References (III)

[Micchelli, Morales, Pontil] **A family of penalty functions for structured sparsity.** NIPS 2010.

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

[Romera-Paredes, Argyriou, Pontil, Berthouze] **Exploiting unrelated tasks in multi-task learning.** AISTATS 2012.

[Salakhutdinov, Torralba, Tenenbaum] **Learning to share visual appearance for multiclass object detection**. CVPR 2011.

[Srivastava and Dwivedi] **Estimation of seemingly unrelated regression equations: A brief survey** J. Econometrics,1971.

[Silver & Mercer] **The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness.** Connection Science 1996. [Yu, Tresp, Schwaighofer] **Learning Gaussian processes from multiple tasks.** ICML 2005.

[Srebro, Rennie, Jaakkola] **Maximum-margin matrix factorization.** NIPS 2004.

[Thrun and Pratt] **Learning to learn**, Springer, 1998.

[Thrun and OSullivan]. **Clustering learning tasks and the selective crosstask transfer of knowledge.** 1998.

[Torralba, Murphy, Freeman] **Sharing features: efficient boosting procedures for multiclass object detection.** CVPR 2004.

[Zellner] **An efficient method for estimating seemingly unrelated regression equations and tests for aggregation bias.** JASA, 1962.