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
- Classes of regularizers
- Statistical analysis
- Optimization methods
- Multilinear models
- Sparse coding

Problem formulation

- Let $\mu_1, \ldots, \mu_T$ be prescribed probability distributions on $X \times Y$
- Goal: find functions $f_t : X \to Y$ which minimize

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

- Learning from data $(x_t^1, y_t^1), \ldots, (x_t^n, y_t^n) \sim \mu_t$, $t = 1, \ldots, T$:

$$\min_{f_1,\ldots,f_T} \left\{ \frac{1}{T} \sum_{t=1}^{T} \frac{1}{n} \sum_{i=1}^{n} L(y_t^i, f_t(x_t^i)) + \lambda\, \Omega(f_1, \ldots, f_T) \right\}$$

- Penalty Ω encourages "common structure" among the functions
- Focus on linear regression: $X \subseteq \mathbb{R}^d$, $Y \subseteq \mathbb{R}$ and $f(x) = \langle w, x \rangle$

Problem formulation (cont.)

- Linear regression model: $y_t^i = \langle w_t^*, x_t^i \rangle + \epsilon_{ti}$

$$\min_{w_1,\ldots,w_T} \; \frac{1}{T} \sum_{t=1}^{T} \underbrace{\frac{1}{n} \sum_{i=1}^{n} L(y_t^i, \langle w_t, x_t^i \rangle)}_{\text{training error task t}} + \lambda \underbrace{\Omega(w_1, \ldots, w_T)}_{\text{joint regularizer}}$$

- Single task learning: $\Omega(w_1, \ldots, w_T) = \sum_t \Omega_t(w_t)$
- 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*

Examples

1. 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, e.g. small variance of parameters
   - special case (matrix completion): $X = \{e_1, \ldots, e_d\}$

2. Multiple object detection in scenes
   - detection of each object corresponds to a binary classification task
   - learning common features enhances performance [Torralba et al. 2004,...]
   - early work in ML using neural nets with shared hidden weights [Baxter 1996, Caruana 1997, Silver and Mercer 1996,...]

More applications in bioinformatics, finance, neuroimaging, NLP, etc.

Quadratic regularizer

$$\min_{w_1,\ldots,w_T} \sum_{t,i} L(y_t^i, \langle w_t, x_t^i \rangle) + \lambda\, \Omega(w_1,\ldots,w_T)$$

- Let $\Omega(w) = \langle w, Ew \rangle$, with $w := (w_1,\ldots,w_T) \in \mathbb{R}^{dT}$ and $E \succ 0$
- *Independent task learning* if E is block diagonal
- Encourage linear relationships between tasks, e.g. [Evgeniou & P., 2004]:

$$\Omega(w) = \sum_{t=1}^{T} \|w_t\|^2 + \frac{1-c}{c} \sum_{t=1}^{T} \left\| w_t - \frac{1}{T} \sum_{s=1}^{T} w_s \right\|^2, \quad c \in [0,1]$$

$c = 1$: independent tasks; $c = 0$: identical tasks

Equivalent formulation
[Evgeniou et al. 2005]

- Choose $v \in \mathbb{R}^p$, $B_t \in \mathbb{R}^{p \times d}$ and set $w_t = {B_t}^\top v$
- Equivalent problem:

$$\min_v \sum_{t,i} L(y_t^i, \langle v, B_t x_t^i \rangle) + \lambda \langle v, v \rangle$$

- Previous ex.: ${B_t}^\top = [(1-c)^{\frac{1}{2}} \mathbf{I}_{d \times d}, \underbrace{\mathbf{0}_{d \times d}, \ldots, \mathbf{0}_{d \times d}}_{t-1}, (cT)^{\frac{1}{2}} \mathbf{I}_{d \times d}, \underbrace{\mathbf{0}_{d \times d}, \ldots, \mathbf{0}_{d \times d}}_{T-t}]$

$$w_t = (1-c)^{\frac{1}{2}} v_0 + (cT)^{\frac{1}{2}} v_t = \text{“}\textit{common}\text{”} + \text{“}\textit{task specific}\text{”}$$

- Extension to coupled hierarchical structures: $w_t = \sum_{k \in A(t)} v_k$
- Connection to kernel methods: learn a map $(x, t) \mapsto f_t(x)$ using the kernel $K((x_1, t_1), (x_2, t_2)) = \langle B_{t_1} x_1, B_{t_2} x_2 \rangle$

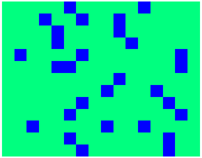
Structured sparsity: few shared variables

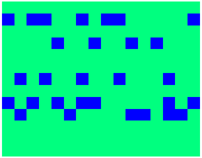
[Argyriou et al. 2006; Lounici et al. 2009]

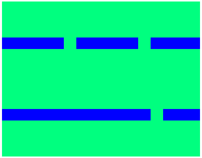
- Favour matrices with many zero rows:

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

Compare matrices W favoured by different regularizers (`green` $= 0$, `blue` $= 1$):







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

Statistical analysis

- Linear regression model: $y_t^i = \langle w_t, x_t^i \rangle + \epsilon_t^i$, with ϵ_t^i 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 $\mathrm{card}\left\{ j \,:\, \sum_{t=1}^{T} w_{tj}^2 > 0 \right\} \leqslant s$
- Variables not too correlated: $\frac{1}{n} \left| \sum_{i=1}^{n} x_{tj}^i x_{tk}^i \right| \leq \frac{1-\rho}{7s}$, $\forall t$, $\forall j \neq k$

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.

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

- 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. 2006, 2008]

Extend above formulation to learn a low dimensional representation:

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

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

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

Equivalent to trace norm regularization:

$$\min_W \sum_{t,i} L(y_t^i, \langle w_t, x_t^i \rangle) + \lambda \, \|W\|_{\mathrm{tr}}$$

Variational form and alternate minimization

- **Fact**: $\|W\|_{\mathrm{tr}} = \frac{1}{2} \inf_{D \succ 0} \mathrm{tr}(D^{-1} W W^\top + D)$ and infimizer $= \sqrt{W W^\top}$

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

- Requires a perturbation step to ensure convergence
- See [Dudík et al. 2012] for comparative results
- Diagonal constraints: $\|W\|_{2,1} = \frac{1}{2} \inf_{z > 0} \left\{ \sum_{j=1}^{d} \frac{\|w_{:,j}\|^2}{z_j} + z_j \right\}$
- Further constraints on z [Micchelli et al. 2010] e.g. ordered structures

Risk bound

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

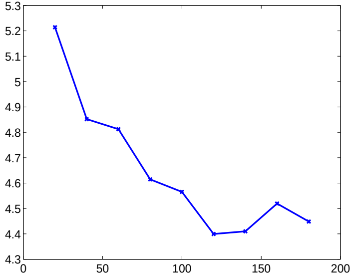
$$R(\hat{W}) - R\left(W^*\right) \leq 2\phi B\left(\sqrt{\frac{\|\hat{C}\|_\infty}{n}} + \sqrt{\frac{2\left(\ln\left(nT\right)+1\right)}{nT}}\right) + \sqrt{\frac{8\ln\left(3/\delta\right)}{nT}}$$

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

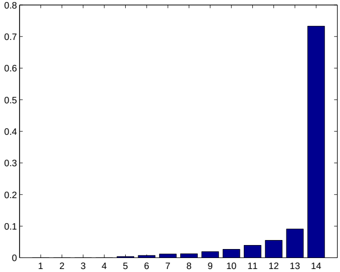
- **Interpretation:** Assume $\operatorname{rank}(W^*) = K$, $\|w_t^*\| \leq 1$ and let $B = \sqrt{K}$. If the inputs are uniformly distributed, as T grows we have a $O(\sqrt{K/nd})$ bound as compared to $O(\sqrt{1/n})$ for single task learning

Experiment (computer survey)

Test error vs. #tasks



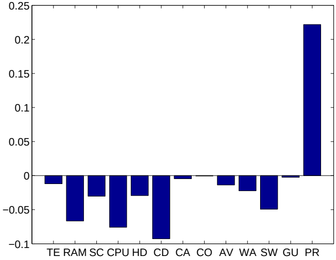
Eigenvalues of D



- 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 $c \in [0, 1]$)	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

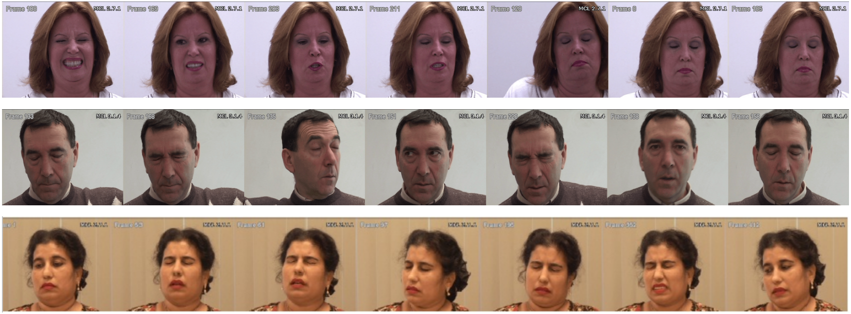
Several possible extension of the above formulations:

- Multiple low dimensional subspaces [Argyriou et al. 08b, Kang et al. 2011,]
- Encourage heterogeneous features [Romera-Paredes et al. 2012]
- Sparse coding [Kumar and Daumé III, 2012, Maurer et al. 2013]
- Multilinear models [Romera-Paredes et al. 2013]

Multilinear MTL

[Romera-Paredes et al. 2013]

- Tasks associated with multi-index, e.g. $t = (t_1, t_2)$
- Example: predict action-units' (e.g. `cheek raiser`) activation for different people [Lucey et. al 2011]: $t_1 \leftrightarrow$ "identity", $t_2 \leftrightarrow$ "action-unit"



Multilinear MTL (cont.)

Let $\mathbf{W} \in \mathbb{R}^{T_1 \times T_2 \times d}$, with $W_{t_1,t_2,:} \in \mathbb{R}^d$ the (t_1, t_2)-th regression task for $t_1 = 1, ..., T_1,\ t_2 = 1, ..., T_2$

- 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]

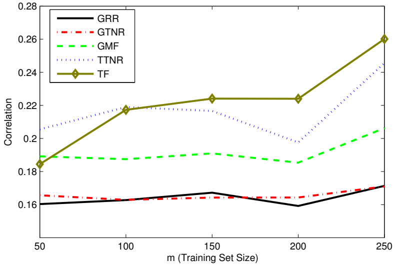
$$\sum_{n=1}^{3} \|W_{(n)}\|_{\mathrm{tr}}$$

- Regularization problem solved by alternating direction of multipliers [Gandy et al. 2011]

- Alternative approach using Tucker decomposition

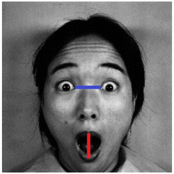
$$W_{t_1,t_2,j} = \sum_{s_1}^{S_1} \sum_{s_2=1}^{S_2} \sum_{k=1}^{K} G_{s_1,s_2,k} A_{t_1,s_1} B_{t_2,s_2} C_{j,k}$$

- Transfer learning experiment (more general setting involving distinct sets of **training** and **target tasks**)



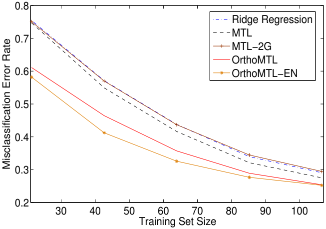
Example: recognizing identity and emotion on a set of faces

- emotion related feature (red)
- identity related feature (blue)



Assumption:

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



$$\min_{W,V} \left\{ \hat{R}_{\rm em}(W) + \hat{R}_{\rm id}(V) + \lambda \|[W, V]\|_{\rm tr} + \rho \|W^\top V\|_{\rm Fr}^2 \right\}$$

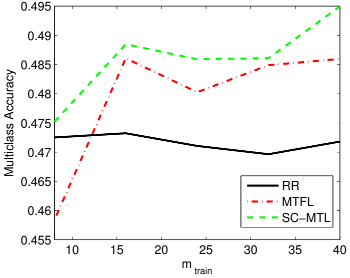
- Related convex problem under conditions

Multi-task learning with dictionaries

[Maurer P., Romera-Paredes, 2013]

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

$$\min_{U,A} \left\{ \sum_{t,i} L(y_t^i, \langle w_t, x_t^i \rangle) : w_t = Ua_t, \|u_k\|_2 \leq 1, \ \|a_t\|_1 \leq \alpha \right\}$$



- Related method with Frobenius norm bound [Kumar and Daumé III, 2012]
- Estimation bounds indicate potential improvement over single task learning

Conclusions

- Multi-task learning is ubiquitous – exploiting task relatedness provides substantial improvement over independent task learning
- Presented families of regularizers which naturally extend complexity notions (smoothness and sparsity) used for single-task learning; touched upon statistical analyses and optimisation methods
- Need for scalable algorithms, particularly in more complex task relatedness scenarios such as multilinear models
- 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

[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, 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.

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

[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.

[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.

[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.

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

[Maurer and Pontil] **K-dimensional coding schemes in Hilbert spaces** IEEE Transactions on Information Theory, 56(11): 5839-5846, 2010.

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

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

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

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

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

[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.

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

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