Multi-task and Transfer Learning

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 and optimization methods
- Sparse coding
- Multilinear models

Problem Formulation

- Fix probability distributions $\mu_1, \ldots, \mu_T$ on $\mathbb{R}^d \times \mathbb{R}$
- Draw data: $\mathbf{z}_t = \big((x_t^1, y_t^1), \ldots, (x_t^m, y_t^m)\big) \sim \mu_t^m, \;\; t = 1,\ldots,T$
- Learn weight vectors $w_1, \ldots, w_T$ by solving

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

Problem Formulation (cont.)

$$\min_{w_1,\ldots,w_T} \ \frac{1}{T}\sum_{t=1}^{T} R(w_t;\mathbf{z}_t) + \lambda\Omega(w_1,\ldots,w_T)$$

- Independent task learning (ITL): $\Omega(w_1,\ldots,w_T) = \sum_t \Omega_t(w_t)$
- Typical scenario: **many tasks** but only **few examples per task**

  In this regime ITL does not work! see [Maurer & Pontil 2008]
- Matrix regularization problem

Applications

- **User modelling:**
  - ◇ each task is to predict a user's ratings to products
  - ◇ 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: $y_t^i \in \{-1, 1\}$
  - ◇ learning common features enhances performance

Many more: affective computing, bioinformatics, neuroimaging, NLP,...

Modelling Task Relatedness – Different Perspectives

Ideas from *kernel methods, sparse estimation, unsupervised learning*

- RKHS of vector-valued functions [Caponnetto et al. 2008, Caponnetto and De Vito 2007, Dinuzzo & Fukumizu 2012, Micchelli and Pontil 2005]
- Variance regularizer [Evgeniou and Pontil 2004, Maurer 2006]
- Common sparsity pattern [Argyriou et al. 2008, Obozinski 2009,..]
- Shared low dimensional subspace (PCA) [Ando and Zhang, 2005, Argyriou et al. 2008,...]
- Multiple low dimensional subspaces [Argyriou et al. 2008a,...]
- Orthogonal tasks [Romera-Paredes et al. 2013a]
- Task clustering [Evgeniou et al. 2005; Jacob et al., 2008]
- Hierarchical relationships [Mroueh et al. 2011; Salakhutdinov et al. 2011]
- Sparse coding [Maurer et al. 2013]

Early work in ML: Use a hidden layer neural network with few nodes and a set of network weights shared by all the tasks [Baxter 1995, Caruana 1994, Silver and Mercer 1996, Thrun and Pratt, 1998,...]

Bayesian approaches: [Archambeau et al. 2011, Bakker & Heskes 2003; Evgeniou et al. 2007; Lenk et al. 96, Xue et al. 2007; Yu et al. 05; Zhang et al. 2006,...]: prior distribution of tasks' parameters

Related areas: conjoint analysis, longitudinal data analysis, seemingly unrelated regression in econometrics, functional data analysis

Examples of Regularizers

- Quadratic, e.g. $\sum_{t=1}^{T} \|w_t - \bar{w}\|_2^2$ or $\sum_{s,t=1}^{T} A_{st} \|w_t - w_s\|_2^2, \quad A_{st} \geq 0$

- Joint sparsity: $\sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{jt}^2}$

- Multitask feature learning: $\|[w_1, ..., w_T]\|_{\mathrm{tr}}$

Quadratic Regularizer

$$\min_{w_1,\ldots,w_T} \frac{1}{T} \sum_t R(w_t; \mathbf{z}_t) + \lambda\, \Omega(w_1, \ldots, w_T)$$

- Example: "stay close to the average" [Evgeniou & Pontil 2004]

$$\Omega(w) = \frac{1}{T} \sum_{t=1}^{T} \|w_t\|^2 + \frac{1-\beta}{\beta} \underbrace{\sum_{t=1}^{T} \left\| w_t - \frac{1}{T} \sum_{s=1}^{T} w_s \right\|^2}_{\text{Variance}(w_1,\ldots,w_T)}, \quad \beta \in [0,1]$$

$\beta = 1$: independent tasks; $\beta = 0$: identical tasks

- If each task is a binary SVM: trade-off margin of each task SVM with variance of the parameters

Quadratic Regularizer (cont.)

Equivalent problem

$$\min_{u_0, u_1 \ldots, u_T} \frac{1}{T} \sum_t R(u_0 + u_t; \mathbf{z}_t) + \lambda \left(\frac{1}{1-\beta} \|u_0\|^2 + \frac{1}{\beta T} \sum_t \|u_t\|^2 \right)$$

To see this:

- Make change of variable: $w_t = u_0 + u_t$
- Minimize over u_0 and use $\mathrm{Variance} = \frac{1}{T} \sum_t \|w_t\|^2 - \|\frac{1}{T} \sum_t w_t\|^2$

Link to Kernel Methods

- Let B_t be prescribed $p \times d$ matrices (typically $p \gg d$)
- Learn function $(x, t) \mapsto f_t(x)$ using feature map $(x, t) \mapsto B_t x$
- Multi-task kernel: $K((x_1, t_1), (x_2, t_2)) = \langle B_{t_1} x_1, B_{t_2} x_2 \rangle$

$$\min_v \frac{1}{Tm} \sum_{t,i} \ell(y_t^i, \langle v, B_t x_t^i \rangle) + \lambda \langle v, v \rangle$$

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

Equivalent to $\min_{w_1,\ldots,w_T} \frac{1}{T} \sum_t R(w_t; \mathbf{z}_t) + \lambda \sum_{s,t=1}^{T} \langle w_s, E_{st} w_t \rangle$

where $E = (B^\top B)^{-1}$, $B = [B_1, ..., B_T]$; see [Evgeniou et al. 2005]

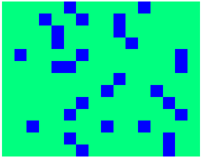
Structured Sparsity: Few Shared Variables

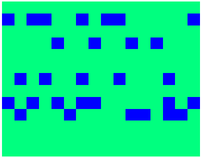
- Favour matrices with many zero rows:

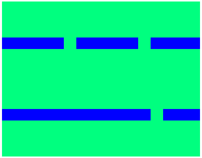
$$\|W\|_{2,1} := \sum_{j=1}^{d} \sqrt{\sum_{t=1}^{T} w_{tj}^2}$$

- Special case of **group Lasso** method [Lounici et al. 09, Obozinski et al. 09, Yuan and Lin 2006]

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: $y_t^i = \langle w_t, x_t^i \rangle + \epsilon_t^i$, with ϵ_t^i i.i.d. $N(0, \sigma^2)$ $i = 1, \ldots, n$, $d \gg n$, use the square loss: $\ell(y, y') = (y - y')^2$
- 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.

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

- Dependency on the dimension d is *negligible* for large T

Multitask Feature Learning

[Argyriou et al. 2008]

Extend above formulation to learn a low dimensional representation:

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

- 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:

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

Variational Form and Alternate Minimization

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

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

- Requires a perturbation step to ensure convergence

- Diagonal constraints: $\|W\|_{2,1} = \frac{1}{2} \inf_{z > 0} \left\{ \sum_{j=1}^{d} \frac{\|w_{:,j}\|^2}{z_j} + z_j \right\}$

- See [Dudík et al. 2012] for comparative results

Risk Bound

Theorem [Maurer & Pontil 2013] Let $R(W) = \frac{1}{T}\sum_{t=1}^{T} \mathbb{E}_{(x,y)\sim\mu_t} \ell(y, \langle w_t, x\rangle)$ and $\hat{R}(W)$ the empirical error. Assume $\ell(y,\cdot)$ is L-Lipschitz and $\|x_t^i\| \leq 1$. If $\hat{W} \in \operatorname{argmin}\{\hat{R}(W) : \|W\|_{tr} \leq B\sqrt{T}\}$ then with probability at least $1-\delta$

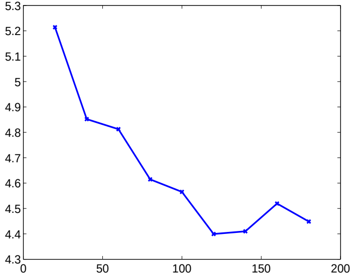
$$R(\hat{W}) - R(W^*) \leq 2LB\left(\sqrt{\frac{\|\hat{C}\|_\infty}{n}} + \sqrt{\frac{2(\ln(nT)+1)}{nT}}\right) + \sqrt{\frac{8\ln(3/\delta)}{nT}}$$

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

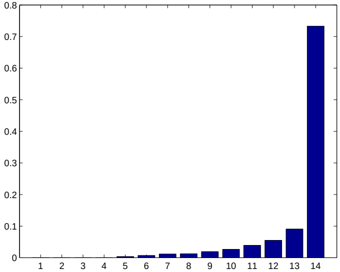
- **Interpretation:** Assume $\operatorname{rank}(W^*) = K$, $\|w_t^*\| \leq 1$ and let $B = \sqrt{K}$. If the inputs are uniformly distributed, as T grows we have a $O(\sqrt{K/nd})$ bound as compared to $O(\sqrt{1/n})$ for single task learning

Experiment (cont.)

Test error vs. #tasks



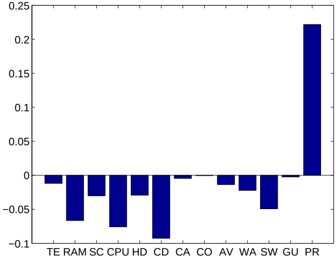
Eigenvalues of D



- 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 (cont.)



Method	Test
Independent	15.05
Aggregate	5.52
Quadratic (best $c \in [0,1]$)	4.37
Structured Sparsity	4.04
MTFL	3.72
Quadratic + Trace	3.20

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

Richer Models of Task Relatedness

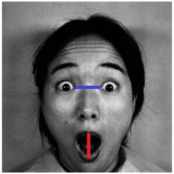
- **Exploiting unrelated tasks / learning heterogeneous features**
- Sparse coding for multitask and transfer learning
- Multilinear models and low rank tensor learning

Exploiting Unrelated Groups of Tasks

[Romera-Paredes et al. 2012]

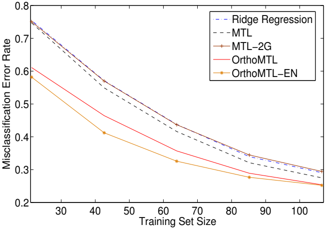
Example: recognizing identity and emotion on a set of faces

- emotion related feature
- identity related feature



Assumption:

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



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

- Related convex problem under conditions

- Exploiting unrelated tasks / encourage heterogeneous features
- **Sparse coding for multitask and transfer learning**
- Multilinear models and low rank tensor learning

Learning Sparse Representations

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

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

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

- Learning method [Maurer et al. 2013]

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

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

Connection to Sparse Coding

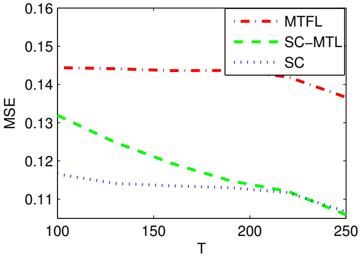
$$\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\|\gamma\|_1 \le \alpha} \frac{1}{m} \sum_{i=1}^{m} \ell\big(\langle D\gamma, x_t^i \rangle, y_t^i\big)$$

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

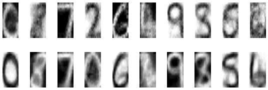
$$\min_{D \in \mathcal{D}_K} \frac{1}{T} \sum_{t=1}^{T} \min_{\|\gamma\|_1 \le \alpha} \|w_t - D\gamma\|_2^2$$

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

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):



Learning Bound

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

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

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

- If T grows, bounds is **comparable to Lasso** with best a-priori known dictionary! [Kakade et al. 2012]

Analysis of Learning to Learn

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

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 & Pontil 2010]

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

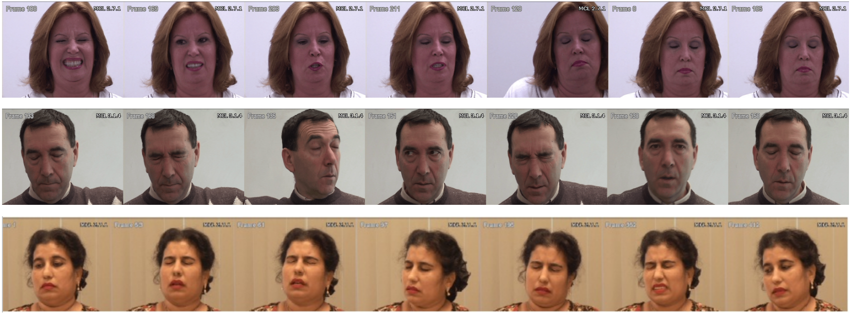
- Exploiting unrelated tasks / encourage heterogeneous features

- Sparse coding for multitask and transfer learning

- **Multilinear models and low rank tensor learning**

Multilinear MTL

- 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, $t_1 = 1,\dots,T_1$, $t_2 = 1,\dots,T_2$
- Goal: control rank of each *matricization* of W:

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

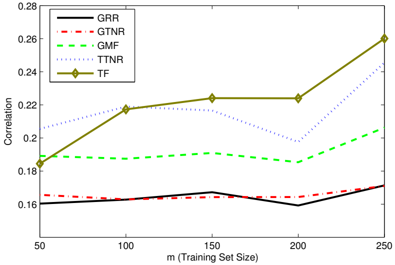
- Convex relaxation [Liu et al. 2009, Gandy et al. 2011, Signoretto et al. 2013]

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

- 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

- $\|\cdot\|_{\text{tr}}$ is the tightest convex relaxation of rank on the spectral unit ball [Fazel et al. 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 & Pontil 2013]

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

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

Related work by [Argyriou et al. 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 $\mathcal{W}$ satisfies conditions below hold 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. 2013]

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

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

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

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 $A = (\boldsymbol{\mathcal{W}} - \frac{1}{\beta}\boldsymbol{\mathcal{C}}_n)_{(n)}$. 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: $\Psi(B) = \psi(\sigma(B))$

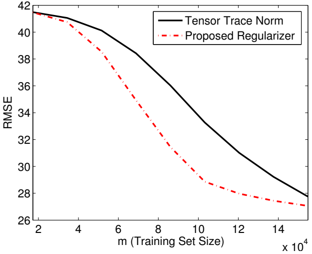
By von Neuman's inequality:

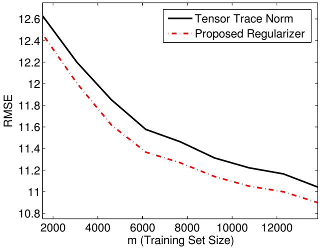
$$\mathrm{prox}_{\frac{1}{\beta}\Psi}(A) = U_A \mathrm{diag}\Big(\mathrm{prox}_{\frac{1}{\beta}\psi}(\sigma_{\mathrm{A}})\Big) V_A^\top$$

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

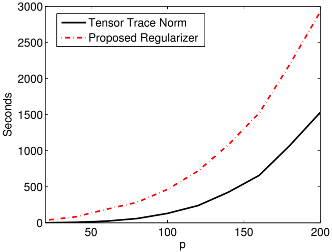
$$\omega_\alpha^*(z) = \sup_{\|x\|_2 \le \alpha} \{\langle x, z\rangle - \mathrm{card}(x)\} = \max_{0 \le r \le d} (\alpha \|z_{1:r}^{\downarrow}\|_2 - r)$$

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
- Reviewed families of regularizers which naturally extend complexity notions (smoothness and sparsity) used for single-task learning
- Recent work on 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

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

[Abernethy et al. 2009] J. Abernethy, F. Bach, T. Evgeniou, J.-P. Vert. A new approach to collaborative filtering: operator estimation with spectral regularization. *JMLR*, 10:803-826, 2009.

[Allenby et al. 1999] Rossi G. M. Allenby and P. E. Rossi. Marketing models of consumer heterogeneity. *Journal of Econometrics*, 89:57–78, 1999.

[Ando & Zhang 2005] R.K. Ando and T. Zhang. A framework for learning predictive structures from multiple tasks and unlabeled data. *JMLR*, 6:1817-1853, 2005.

[Archambeau et al. 2005] Sparse bayesian multi-task learning. C. Archambeau, S. Guo, O Zoeter. Advances in Neural Information Processing Systems 24:1755-1763, 2011.

[Argyriou et al. 2009] A. Argyriou, C.A. Micchelli, M. Pontil. When is there a representer theorem? Vector versus matrix regularizers. *JMLR*, 10:2507-2529, 2009.

[Argyriou et al. 2008] A. Argyriou, T. Evgeniou, M. Pontil. Convex multi-task feature learning. *Machine Learning*, 73(3):243–272, 2008.

[Argyriou et al. 2008a] A. Argyriou, A. Maurer, M. Pontil. An algorithm for transfer learning in a heterogeneous environment. *ECML* 2008.

[Argyriou et al. 2012] A. Argyriou, R. Foygel and N. Srebro. Sparse Prediction with the k-Support Norm. *NIPS*, 2012.

[Argyriou et al. 2011] A. Argyriou, C.A. Micchelli, M. Pontil, L. Shen and Y. Xu. Efficient first order methods for linear composite regularizers. *arXiv:1104.1436*, 2011.

[Arora et al. 1998] N. Arora G.M Allenby, and J. Ginter. A hierarchical Bayes model of primary and secondary demand. *Marketing Science*, 17(1):29–44, 1998.

[Bakker & Heskes 2003] B. Bakker and T. Heskes. Task clustering and gating for Bayesian multitask learning. *JMLR*, 2003.

[Baldassarre et al. 2013] L. Baldassarre, L. Rosasco, A. Barla and A. Verri, Multi-Output Learning via Spectral Filtering. *Machine Learning*, 83(3), 2013.

[Baxter 2000] J. Baxter. A model for inductive bias learning. *J. Artificial Intelligence Research*, 12:149–198, 2000.

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

[Ben-David et al. 2002] S. Ben-David, J. Gehrke, and R. Schuller. A theoretical framework for learning from a pool of disparate data sources. *Proc. of Knowledge Discovery and Datamining (KDD)*, 2002.

[Bertsekas & Tsitsiklis 1989] D.P. Bertsekas, J.N. Tsitsiklis. *Parallel and Distributed Computation: Numerical Methods*. Prentice-Hall, 1989.

[Boyd et al. 2003] S. Boyd, L. Xiao, A. Mutapcic. Subgradient methods, Stanford University, 2003.

[Combettes & Pesquet 2011] P. L. Combettes and J.-C. Pesquet. Proximal splitting methods in signal processing. In *Fixed-Point Algorithms for Inverse Problems in Science and Engineering* (H. H. Bauschke et al. Eds), pages 185-212, Springer, 2011.

[Caponetto et al. 2008] CMPY A. Caponnetto, C.A. Micchelli, M. Pontil, Y. Ying. Universal mult-task kernels. *JMLR* 2008.

[Caponnetto & De Vito 2007] A. Caponnetto, E. De Vito. Optimal rates for regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331-368, 2007.

[Caruana 1998] R. Caruana. Multi-task learning. *Machine Learning*, 1998.

[Cavallanti et al. 2010] G. Cavallanti, N. Cesa-Bianchi, C. Gentile. Linear algorithms for online multitask classification, *JMLR* 2010.

[Dinuzzo & Fukumizu 2011] F. Dinuzzo and K. Fukumizu. Learning low-rank output kernels. ACML 2011.

[Dudík et al. 2012] M. Dudík, Z. Harchaoui, J. Malik. Lifted coordinate descent for learning with trace-norm regularization, Proc. AISTATS, 2012.

[Evgeniou & Pontil 2004] T. Evgeniou and M. Pontil. Regularized multi-task learning. *SIGKDD*, 2004.

[Evgeniou et al. 2007] T. Evgeniou, M. Pontil, O. Toubia. A convex optimization approach to modeling heterogeneity in conjoint estimation. *Marketing Science*, 26:805-818, 2007.

[Evgeniou et al. 2005] T. Evgeniou, C.A. Micchelli, M. Pontil. Learning multiple tasks with kernel methods. *JMLR* 2005.

[Fazel et al. 2001] M. Fazel, H. Hindi, and S. Boyd. A rank minimization heuristic with application to minimum order system approximation. *Proc. American Control Conference*, Vol. 6, pages 4734-4739, 2001.

[Gandy et al. 2011] S. Gandy, B. Recht, I. Yamada. Tensor completion and low-n-rank tensor recovery via convex optimization. *Inverse Problems*, 27(2), 2011.

[Hiriart-Urruty et al. 1996] J-B. Hiriart-Urruty and C. Lemaréchal. *Convex Analysis and Minimization Algorithms* Springer, 1996.

[Izenman 1975] A. J. Izenman. Reduced-rank regression for the multivariate linear model. *Journal of Multivariate Analysis*, 5:248-264, 1975.

[Jebara 2004] T. Jebara. Multi-task feature and kernel selection for SVMs. *Proc. 21st International Conference on Machine Learning*, 2004.

[Jebara 2011] T. Jebara. Multitask sparsity via maximum entropy discrimination. *JMLR*, 12:75-110, 2011.

[Jacob et al. 2008] L. Jacob, F. Bach, J.-P. Vert. Clustered multi-task learning: a convex formulation. NIPS 2008.

[Kakade et al. 2012] S. Kakade, S. Shalev-Shwartz, A.Tewari. Regularization techniques for learning with matrices. *JMLR*, 2012.

[Kang et al. 2011] Z. Kang, K. Grauman, F. Sha. Learning with whom to share in multi-task feature learning. *Proc. ICML*, 2011.

[Kolda & Bader 2009] T.G. Kolda and B.W. Bader. Tensor decompositions and applications. *SIAM Review*, 51(3):455-500, 2009.

[Kumar & Daumé III 2012] A. Kumar and H. Daumé III. Learning task grouping and overlap in multi-task learning. *Proc. ICML*, 2012.

[Lawrence & Platt 2004] N. D. Lawrence and J. C. Platt. Learning to learn with the informative vector machine. *Proc. 21-st International Conference in Machine Learning*, 2004.

[Lenk et al. 1996] P. J. Lenk, W. S. DeSarbo, P. E. Green, and M. R. Young. Hierarchical Bayes conjoint analysis: recovery of partworth heterogeneity from reduced experimental designs. *Marketing Science*, 15(2):173-191, 1996.

[Liu et al. 2009] J. Liu, P. Musialski, P. Wonka, J. Ye. Tensor completion for estimating missing values in visual data. *Proc. 12th International Conference on Computer Vision (ICCV)*, pages 2114-2121, 2009.

[Lounici et al. 2011] K. Lounici, M. Pontil, A. Tsybakov, S. van de Geer. Oracle inequalities and optimal inference under group sparsity. *Annals of Stat.*, 2011.

[Lucey et al. 2011] P. Lucey, J.F. Cohn, K.M. Prkachin, P.E. Solomon, I. Matthews. Painful data: The UNBC-McMaster shoulder pain expression archive database. Proc. Int. Conf. Automatic Face and Gesture Recognition, pages 57-64, 2011.

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

[Maurer & Pontil 2008] A. Maurer and M. Pontil. A uniform lower error bound for half-space learning. *Proc. ALT*, 2008.

[Maurer & Pontil 2010] A. Maurer and M. Pontil. K-dimensional coding schemes in Hilbert spaces. *IEEE Trans. Information Theory*, 2010.

[Maurer & Pontil 2013] A. Maurer and M. Pontil. Excess risk bounds for multitask learning with trace norm regularization . *Proc. COLT*, 2013.

[Maurer et al. 2013] A. Maurer, M. Pontil, B. Romera-Paredes. Sparse coding for multitask and transfer learning. *Proc. ICML* 2013.

[Micchelli et al. 2013] C.A. Micchelli, J. Morales, M. Pontil. Regularizers for structured sparsity. *Advances in Computational Mathematics*, 38(3):455-489, 2013.

[Micchelli & Pontil 2005] C.A. Micchelli and M. Pontil. On learning vector-valued functions. *Neural Computation*, 2005.

[Mroueh et al. 2011] Y. Mroueh, T. Poggio, L. Rosasco. Regularization Predicts While Discovering Taxonomy. *Technical Report*, Massachusetts Institute of Technology, 2011.

[Nesterov 2007] Y. Nesterov. Gradient methods for minimizing composite objective functions. *ECORE Discussion Paper*, 2007/96, 2007.

[Rasmussen & Williams] C.E. Rasmussen and C. Williams. *Gaussian Processes for Machine Learning*, MIT Press, 2006.

[Romera-Paredes et al. 2012] B. Romera-Paredes, A. Argyriou, N. Bianchi-Berthouze, M. Pontil. Exploiting unrelated tasks in multi-task learning. *Proc. AISTATS*, 2012.

[Romera-Paredes et al. 2013a] B. Romera-Paredes, H. Aung, N. Bianchi-Berthouze and M. Pontil. Multilinear multitask learning. *Proc. 30th International Conference on Machine Learning (ICML)*, pages 1444-1452, 2013.

[Romera-Paredes et al. 2013b] B. Romera-Paredes, H. Aung, M. Pontil, A.C. Williams, P. Watson, N. Bianchi-Berthouze. Transfer learning to account for idiosyncrasy in face and body expressions*Proc. 10th International Conference on Automatic Face and Gesture Recognition (FG)*, 2013.

[Romera-Paredes & Pontil 2013] B. Romera-Paredes and M. Pontil. A new convex relaxation for tensor completion. *arXiv:1307.4653*, 2013.

[Shor 1985] N. Z. Shor. *Minimization Methods for Non-differentiable Functions*. Springer, 1985.

[Signoretto et al. 2013] M. Signoretto, Q. Tran Dinh, L. De Lathauwer, J.A.K. Suykens. Learning with tensors: a framework based on convex optimization and spectral regularization. *Machine Learning*, to appear.

[Signoretto et al. 2011] M. Signoretto, R. Van de Plas, B. De Moor, J.A.K. Suykens. Tensor versus matrix completion: a comparison with application to spectral data. *IEEE Signal Processing Letters*, 18(7):403-406, 2011.

[Silver & Mercer 1996] D. L. Silver and R.E Mercer. The parallel transfer of task knowledge using dynamic learning rates based on a measure of relatedness. *Connection Science*, 8:277-294, 1996.

[Srebro et al. 2005] N. Srebro, J. Rennie and T. Jaakkola. Maximum margin matrix factorization. *Advances in Neural Information Processing Systems (NIPS) 17*, pages 1329-1336, 2005.

[Salakhutdinov et al. 2011] R. Salakhutdinov, A. Torralba, J. Tenenbaum. Learning to Share Visual Appearance for Multiclass Object Detection *Proc. CVPR*, 2011.

[Tomioka et al. 2010] R. Tomioka, K. Hayashi, H. Kashima, J.S.T. Presto. Estimation of low-rank tensors via convex optimization. *arXiv:1010.0789*, 2010.

[Tomioka et al. 2013] R. Tomioka, T. Suzuki, K. Hayashi, H. Kashima. Statistical performance of convex tensor decomposition. *Advances in Neural Information Processing Systems (NIPS) 24*, pages 972-980, 2013.

[Thrun & Pratt 1998] S. Thrun and L.Y. Pratt. *Learning to Learn*, Springer, 1998.

[Thrun & O'Sullivan 1998] Thrun and OSullivan. Clustering learning tasks and the selective crosstask transfer of knowledge. 1998.

[Torralba et al. 2004] A. Torralba, K.P. Murphy, W.T. Freeman. Sharing features: efficient boosting procedures for multiclass object detection. *Proc. CVPR*, 2004.

[Yu et al. 2005] K. Yu, V. Tresp, and A. Schwaighofer. Learning Gaussian processes from multiple tasks.

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

[Zhang et al. 2006] Zhang, Z. Ghahramani, and Y. Yang. Learning multiple related tasks. using latent independent component analysis. In Advances in Neural Information Processing Systems 18, pages 1585-1592, 2006.