BOUNDS FOR LINEAR MTL

Andreas Maurer

ingredients of linear MTL

- independent random variables $\left(X^1, Y^1\right), ..., (X^m, Y^m) : \Omega \to H \times \mathbb{R}$ where H is a Hilbert-space and $\left\|X^l\right\| \leq 1$. Here m is the number of tasks.

- loss functions $\phi^1, ..., \phi^m : \mathbb{R} \times \mathbb{R} \to [0, 1]$ where $\phi^l(y, .)$ is 1-Lipschitz.

- training sample $(\mathbf{X}, \mathbf{Y}) = \left(\left(X_i^l, Y_i^l\right) : 1 \leq l \leq m, 1 \leq i \leq n\right)$ with $\left(X_i^l, Y_i^l\right)_{i=1}^n$ indep., identically distributed to $\left(X^l, Y^l\right)$.

- a hypothesis class $\mathcal{F}$ of linear transformations $V : H \to \mathbb{R}^m$, where the l-th row-vector v^l of V is the linear predictor for the l-th task

$$(Vx)_l = \left\langle v^l, x\right\rangle.$$

objective

Find a transformation $V = \left(v^1, ..., v^m\right) \in \mathcal{F}$ with small task-averaged loss

$$\text{er}\left(V\right) = \frac{1}{m}\sum_{l=1}^{m}\mathbb{E}\left[\phi^l\left(Y^l, \left\langle v^l, X^l\right\rangle\right)\right].$$

Since the distribution of the $\left(X^l, Y^l\right)$ is unknown, we work on the basis of an empirical estimate

$$\hat{\text{er}}\left(V\right)\left(\mathbf{X}, \mathbf{Y}\right) = \frac{1}{m}\sum_{l=1}^{m}\frac{1}{n}\sum_{i=1}^{n}\phi^l\left(Y_i^l, \left\langle v^l, X_i^l\right\rangle\right).$$

error bound

(Kolchinski+Panchenko, Bartlett+Mendelson, Ando+Zhang)

$\mathcal{F}$ a class of linear transformations $V : H \to \mathbb{R}^m$.

$\forall \delta$ with probability at least $\mathbf{1} - \delta$ in the sample $(\mathbf{X}, \mathbf{Y})$, we have $\forall V \in \mathcal{F}$

$$\mathrm{er}\,(V) \;\le\; \hat{\mathrm{er}}\,(V)\,(\mathbf{X}, \mathbf{Y}) \;+\; \hat{\mathcal{R}}_n^m\,(\mathcal{F})\,(\mathbf{X}) \;+\; \sqrt{\frac{9\ln(2/\delta)}{2mn}}$$

Model selection corollary:

If J and D are such that $\forall \mathcal{F}$ we have $\hat{\mathcal{R}}_n^m\,(\mathcal{F})\,(\mathbf{X}) \le \sup_{V\in\mathcal{F}} J\,(V)\,D\,(\mathbf{X})$ then whp for *all* $V : H \to \mathbb{R}^m$

$$\mathrm{er}\,(V) \;\le\; \hat{\mathrm{er}}\,(V)\,(\mathbf{X}, \mathbf{Y}) \;+\; 2J\,(V)\,D\,(\mathbf{X}) \;+\; \sqrt{\frac{9\ln(2J(V)/\delta)}{2mn}}\;.$$

Multi-task regularization: Impose 'relatedness' or 'similarity' constraints among the task-specific predictors v^l for $V = \left(v^1, ..., v^m\right)$.

Rademacher complexity

$\mathcal{F}$ a class of linear transformations $V : H \to \mathbb{R}^m$.
σ_i^l independent random variables, distributed uniformly in $\{-1, 1\}$.

$$\begin{aligned}\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) &= E_\sigma \left[\sup_{V \in \mathcal{F}} \frac{2}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \sigma_i^l \left\langle v^l, X_i^l \right\rangle \right] \\ &= E_\sigma \left[\sup_{V \in \mathcal{F}} \frac{2}{mn} tr\left(V^* W_{\boldsymbol{\sigma}, \mathbf{X}} \right) \right]\end{aligned}$$

where $W_{\boldsymbol{\sigma}, \mathbf{X}} : H \to R^m$ is the transformation

$$\left(W_{\boldsymbol{\sigma}, \mathbf{X}}\, z \right)_l = \left\langle \sum_{i=1}^{n} \sigma_i^l X_i^l, z \right\rangle .$$

Hölder's inequality

For a compact operator $A : H \to H'$ define $\|A\|_p = \left(\sum_i \mu_i (A)^p\right)^{1/p}$, where μ_i are the singular values of A.

Theorem: For $p^{-1} + q^{-1} = 1$ we have $|tr\,(A^* B)| \leq \|A\|_p \|B\|_q$.

$$\begin{aligned} \text{So} \quad \hat{\mathcal{R}}_n^m (\mathcal{F})(\mathbf{X}) &= E_\sigma \left[\sup_{V \in \mathcal{F}} \frac{2}{mn} tr \left(V^* W_{\boldsymbol{\sigma},\mathbf{X}} \right) \right] \\ &\leq \frac{2}{\sqrt{n}} \left(\sup_{V \in \mathcal{F}} \frac{\|V\|_q}{\sqrt{m}} \right) E_\sigma \left[\frac{\left\| W_{\boldsymbol{\sigma},\mathbf{X}} \right\|_p}{\sqrt{mn}} \right]. \end{aligned}$$

Bound last factor to get

theorem

Let $\mathcal{F}$ be any set of linear transformations $V : H \to \mathbb{R}^m$.
$p \in [4, \infty]$ and $p^{-1} + q^{-1} = 1$.

$$\hat{\mathcal{R}}_n^m (\mathcal{F}) (\mathbf{X}) \leq \frac{2}{\sqrt{n}} \left(\sup_{V \in \mathcal{F}} \frac{\|V\|_q}{\sqrt{m}} \right) \sqrt{\left\| \hat{C} (\mathbf{X}) \right\|_{p/2} + \sqrt{\frac{2}{m}}}$$

where $\hat{C} (\mathbf{X})$ is the *empirical covariance operator*

$$\hat{C} (\mathbf{X}) z = \frac{1}{mn} \sum_{l=1}^{m} \sum_{i=1}^{n} \left\langle z, X_i^l \right\rangle X_i^l \text{ for } z \in H.$$

multi-task subspace learning

$$
\begin{aligned}
\mathcal{F}_{B,d} &= \left\{ V : \frac{1}{m} \sum_l \left\| v^l \right\|^2 \leq B^2 \text{ and } \operatorname{rank}(V) \leq d \right\} \\
\implies & \sup_{V \in \mathcal{F}_{B,d}} \frac{\|V\|_q}{\sqrt{m}} \leq B d^{\frac{p-2}{2p}}
\end{aligned}
$$

For homogeneous data-distribution on k-sphere $\|C\|_{p/2} = k^{\frac{2-p}{p}}$, so

$$
\begin{aligned}
E_{\mathbf{X}} \left[\hat{\mathcal{R}}_n^m \left(\mathcal{F}_{B,d} \right) \mathbf{X} \right] &\leq \frac{2B}{\sqrt{n}} \sqrt{ \left(\frac{d}{k} \right)^{\frac{p-2}{p}} + d^{\frac{p-2}{p}} \sqrt{\frac{3}{m}} } \\
&\to \frac{2B}{\sqrt{n}} \sqrt{\frac{d}{k}} \text{ if } p = \infty \text{ and } m \to \infty.
\end{aligned}
$$

graph regularization

(Evgeniou+Micchelli+Pontil)

Suppose w_{lr} quantifies our belief in the similarity of tasks l and r,
Assume $w_{lr} = w_{rl}$, $w_{lr} \geq 0$ and connectedness.

Suggests regularizer

$$\begin{aligned} J(V)^2 &= \frac{1}{2m} \sum_{l,r} w_{lr} \left\| v^l - v^r \right\|^2 + \frac{\eta}{m} \sum_{l=1}^{m} \left\| v^l \right\|^2 \\ &= \frac{1}{m} tr\left(V^* \left(\Delta + \eta I \right) V \right), \end{aligned}$$

where Δ is the m-vertex graph-Laplacian with edge-weights w.

a bound for graph regularization

With $A := \mathbf{\Delta} + \eta I$ (then A is non-singular and ≥ 0)

$$\begin{aligned}
\hat{\mathcal{R}}_n^m(\mathcal{F})(\mathbf{X}) &= E_\sigma\left[\sup_{V\in\mathcal{F}} \frac{2}{mn} tr\left(V^* A^{1/2} A^{-1/2} W_{\boldsymbol{\sigma},\mathbf{X}}\right)\right] \quad (\ \leftarrow \text{trick}) \\
&\leq \frac{2}{\sqrt{n}} \sup_{V\in\mathcal{F}} \left(\frac{tr\,(V^* A V)}{m}\right)^{1/2} \left(\frac{E_\sigma\left[tr\left(W^*_{\boldsymbol{\sigma},\mathbf{X}} A^{-1} W_{\boldsymbol{\sigma},\mathbf{X}}\right)\right]}{mn}\right)^{1/2} \\
&\leq \frac{2}{\sqrt{n}} \sup_{V\in\mathcal{F}} J(V) \sqrt{\frac{1}{m}\sum_{i=2}^{m} \frac{1}{\lambda_i + \eta} + \frac{1}{m\eta}},
\end{aligned}$$

where λ_i are the eigenvalues of $\mathbf{\Delta}$ in non-decreasing order.