An inequality with applications to structured sparsity and multitask dictionary learning

Andreas Maurer

(joint work with Massimiliano Pontil and Bernardino Romera-Paredes - both UCL London)

Uniform bounds with Rademacher complexities

Theorem (Bartlett, Mendelson, 2002).
$\mathcal{F} : \mathcal{X} \to [0,1]$, $X \sim \mu$, $\mathbf{X} = (X_1, ..., X_n) \sim \mu^n$.

$$\Pr\left\{ \sup_{f \in \mathcal{F}} \left(\mathbb{E} f(X) - \frac{1}{n} \sum f(X_i) \right) > \mathcal{R}(\mathcal{F}, \mathbf{X}) + \sqrt{\frac{9 \ln(2/\delta)}{2n}} \right\} \leq \delta,$$

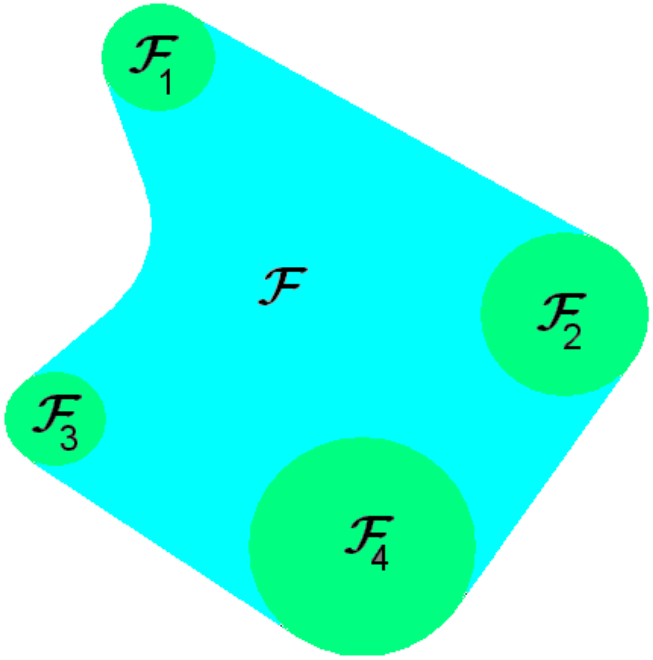
$$\text{with } \mathcal{R}(\mathcal{F}, \mathbf{x}) := \frac{2}{n} \mathbb{E} \sup_{f \in \mathcal{F}} \sum_{i=1}^{n} \epsilon_i f(x_i)$$

with ϵ_i independent, uniform on $\{-1, 1\}$.

- $\mathcal{R}(\phi \circ \mathcal{F}, \mathbf{x}) \leq \|\phi\|_{Lip} \mathcal{R}(\mathcal{F}, \mathbf{x})$
- $\mathcal{R}(\mathcal{F}, \mathbf{x}) = \mathcal{R}(Co(\mathcal{F}), \mathbf{x})$.

Unions of classes

Often there are $\mathcal{F}_1, ..., \mathcal{F}_M$ such that $\mathcal{R}(\mathcal{F}, \mathbf{x}) \leq \mathcal{R}(\bigcup_m \mathcal{F}_m, \mathbf{x})$, for example when $ext(\mathcal{F}) \subseteq \bigcup_m \mathcal{F}_m$.



- Structured sparsity methods (multiple kernel learning, group lasso, K-support norm,...

- Multi-task dictionary learning (multi-task sparse coding, subspace learning,...

Rewriting the Rademacher average of unions

$$\mathcal{R}\left(\bigcup_m \mathcal{F}_m, \mathbf{x}\right) = \frac{2}{n}\mathbb{E}\max_m \left[\sup_{f\in\mathcal{F}_m} \sum_{i=1}^n \epsilon_i f\left(x_i\right)\right] = \frac{2}{n}\mathbb{E}\max_m F_m\left(\boldsymbol{\epsilon}\right)$$

$$\text{with } F_m\left(\mathbf{z}\right) = \sup_{f\in\mathcal{F}_m} \sum_{i=1}^n z_i f\left(x_i\right), \text{ for } \mathbf{z}\in\mathbb{R}^n.$$

- F_m is convex on $\mathbb{R}^n$
- By Cauchy Schwarz F_m is Lipschitz with constant

$$\sqrt{\sup_{f\in\mathcal{F}_m} \sum_{i=1}^n f^2\left(x_i\right)} \leq \sqrt{\sup_{f\in\mathcal{F}} \sum_{i=1}^n f^2\left(x_i\right)} =: L$$

Concentration of convex Lipschitz functions

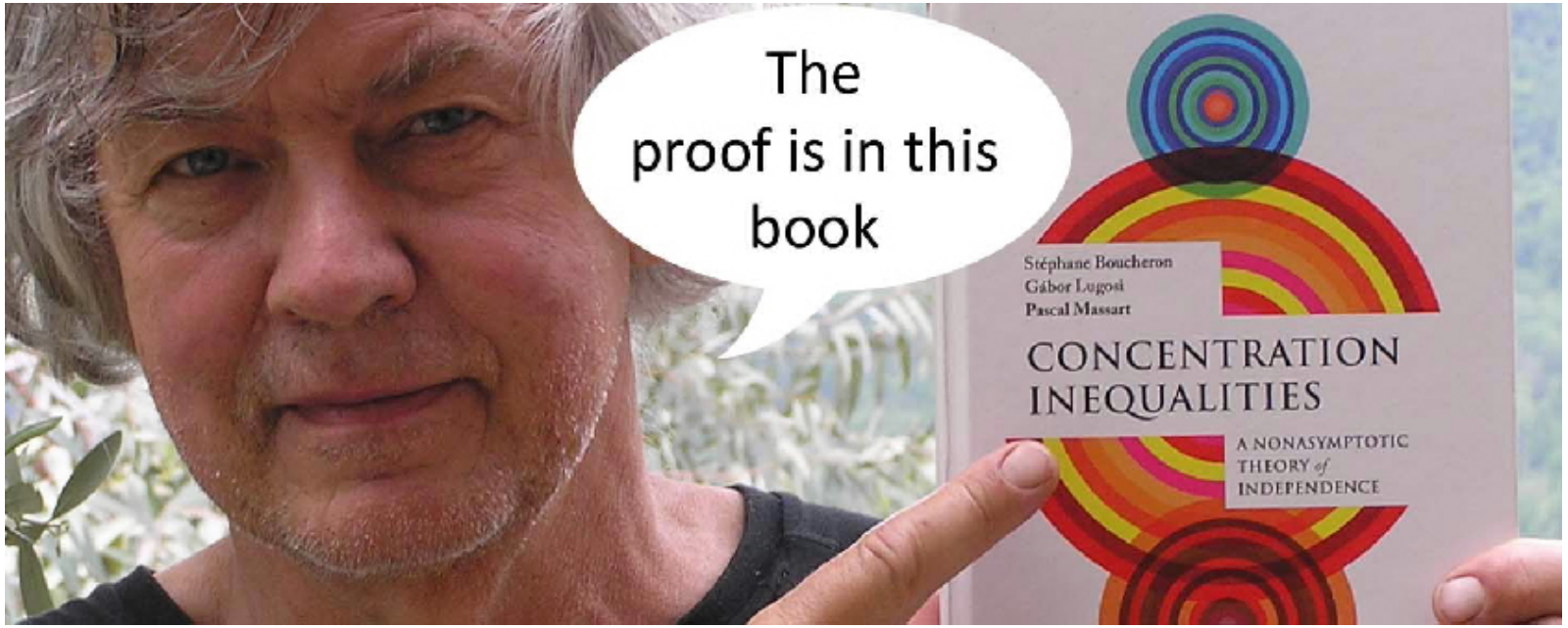
Theorem: If $F : [-1,1]^n \to \mathbb{R}$ is convex and L-Lipschitz and $\mathbf{X} = (X_1, ..., X_n)$ is a vector of independent r.v. with values in $[-1,1]$ then for every $\beta > 0$

$$\mathbb{E} \exp\left(\beta \left[F\left(\mathbf{X}\right) - \mathbb{E}F\left(\mathbf{X}\right)\right]\right) \leq e^{2\beta^2 L^2}.$$

Concentration of convex Lipschitz functions

Theorem: If $F : [-1,1]^n \to \mathbb{R}$ is convex and L-Lipschitz and $\mathbf{X} = (X_1, ..., X_n)$ is a vector of independent r.v. with values in $[-1,1]$ then for every $\beta > 0$

$$\mathbb{E} \exp\left(\beta \left[F\left(\mathbf{X}\right) - \mathbb{E}F\left(\mathbf{X}\right)\right]\right) \leq e^{2\beta^2 L^2}.$$



Bounding the Rademacher average of unions

$$\begin{aligned}
\forall \beta > 0\,, \quad \exp\left(\beta \mathbb{E} \max_m \left(F_m - \mathbb{E}F_m\right)\right) &\leq \mathbb{E} \max_m \exp\left(\beta\left(F_m - \mathbb{E}F_m\right)\right) \\
&\leq \sum_m \mathbb{E} \exp\left(\beta\left(F_m - \mathbb{E}F_m\right)\right) \\
&\leq M e^{2\beta^2 L^2}
\end{aligned}$$

Bounding the Rademacher average of unions

$$\begin{aligned}
\forall \beta > 0\,, \quad \exp\left(\beta \mathbb{E} \max_m \left(F_m - \mathbb{E}F_m\right)\right) &\leq \mathbb{E} \max_m \exp\left(\beta\left(F_m - \mathbb{E}F_m\right)\right) \\
&\leq \sum_m \mathbb{E} \exp\left(\beta\left(F_m - \mathbb{E}F_m\right)\right) \\
&\leq M e^{2\beta^2 L^2}
\end{aligned}$$

Take log, divide by β, use triangle inequality and optimize in β

$$\begin{aligned}
\mathbb{E} \max_m \left(F_m\right) &\leq \max_m \mathbb{E}F_m + \frac{\ln M}{\beta} + 2\beta L^2 \\
&\leq \max_m \mathbb{E}F_m + L\sqrt{8 \ln M}
\end{aligned}$$

Bounding the Rademacher average of unions

$$\begin{aligned}\forall \beta > 0\,, \;\; \exp\left(\beta \mathbb{E} \max_m \left(F_m - \mathbb{E}F_m\right)\right) &\leq \mathbb{E} \max_m \exp\left(\beta\left(F_m - \mathbb{E}F_m\right)\right) \\ &\leq \sum_m \mathbb{E} \exp\left(\beta\left(F_m - \mathbb{E}F_m\right)\right) \\ &\leq M e^{2\beta^2 L^2}\end{aligned}$$

Take log, divide by β, use triangle inequality and optimize in β

$$\begin{aligned}\mathbb{E} \max_m \left(F_m\right) &\leq \max_m \mathbb{E}F_m + \frac{\ln M}{\beta} + 2\beta L^2 \\ &\leq \max_m \mathbb{E}F_m + L\sqrt{8 \ln M}\end{aligned}$$

We have shown:

$$\mathcal{R}\left(\bigcup_m \mathcal{F}_m, \mathbf{x}\right) \leq \max_m \mathcal{R}\left(\mathcal{F}_m, \mathbf{x}\right) + \frac{8}{n}\sqrt{\sup_{f \in \mathcal{F}} \sum_{i=1}^{n} f^2\left(x_i\right) \ln M}$$

Our "main result"

Lemma:

$$\mathcal{R}\left(\bigcup_m \mathcal{F}_m, \mathbf{x}\right) \leq \mathfrak{S} + 8\mathfrak{W}\sqrt{\frac{\ln M}{n}},$$

with $\mathfrak{S} = \max_m \mathcal{R}\left(\mathcal{F}_m, \mathbf{x}\right)$ and $\mathfrak{W} = \max_m \sqrt{\sup_{f \in \mathcal{F}_m} \frac{1}{n} \sum_{i=1}^{n} f^2\left(x_i\right)}$.

$\mathfrak{S}$ = strong parameter, complexity of worst class,

$\mathfrak{W}$ = weak parameter, small for high semantic specificity
(e.g. narrow Gaussian kernels).

Example

$\phi_1, ..., \phi_M$ fixed featuremaps, $\phi_m : \mathcal{X} \to H$,
$\mathcal{F}_m = \{x \mapsto \langle \phi_m(x), w \rangle : \|w\| \leq 1\}$
For $\mathbf{x} \in \mathcal{X}^n$ define $\left\langle \hat{C}_m v, w \right\rangle = (1/n) \sum_i \langle \phi_m(x_i), v \rangle \langle \phi_m(x_i), w \rangle$

$$\mathfrak{S} \leq \max_m \frac{2}{n} \sqrt{\sum_i \|\phi_m(x_i)\|^2} = 2 \max_m \sqrt{\frac{tr\left(\hat{C}_m\right)}{n}}$$

$$\mathfrak{W} = \max_m \sqrt{\sup_{\|w\| \leq 1} \frac{1}{n} \sum_{i=1}^{n} \langle \phi_m(x_i), w \rangle^2} = \max_m \sqrt{\lambda_{\max}\left(\hat{C}_m\right)}$$

Resulting bound:

$$\mathcal{R}\left(\bigcup_m \mathcal{F}_m, \mathbf{x}\right) \leq 2 \max_m \sqrt{\frac{tr\left(\hat{C}_m\right)}{n}} + 8 \max_m \sqrt{\frac{\lambda_{\max}\left(\hat{C}_m\right) \ln M}{n}}$$

Same bound for convex hull (structured sparsity)

$$\mathcal{F} = Co\left(\bigcup_m \mathcal{F}_m\right)$$

is the hypothesis space of

- multiple kernel learning if ϕ_m is the feature-map induced by a p.d. kernel κ on $\mathcal{X} \times \mathcal{X}$
- the group lasso if $\mathcal{X} = \mathbb{R}^d$ and ϕ_m is the projection to $Span\left(e_j : j \in J_m\right)$.
- works also if the J's overlap.
- ...

$$\mathcal{R}\left(Co\left(\bigcup_m \mathcal{F}_m\right), \mathbf{x}\right) \leq 2 \max_m \sqrt{\frac{tr\left(\hat{C}_m\right)}{n}} + 8 \max_m \sqrt{\frac{\lambda_{\max}\left(\hat{C}_m\right) \ln M}{n}}$$

Multitask learning

There are T tasks.

$\mathcal{F}$ a class of vector valued $\mathbf{f} : x \in \mathcal{X} \mapsto \mathbf{f}(x) = (f_1(x), ..., f_1(x)) \in \mathbb{R}^T$.

$(x_{t1}, ..., x_{tn}) =$ examples available for t-th task.

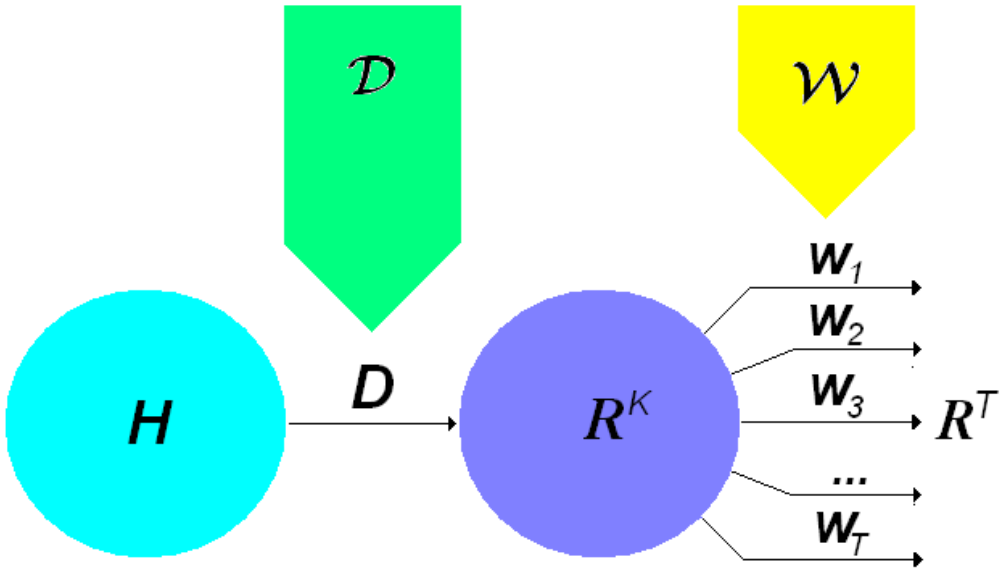
Multitask Rademacher complexity

$$\mathcal{R}(\mathcal{F}, \mathbf{x}) = \frac{2}{nT}\mathbb{E} \sup_{\mathbf{f}\in\mathcal{F}} \sum_{ti} \epsilon_{ti} f_t(x_{ti})$$

used to bound uniform task-average estimation error

$$\sup_{\mathbf{f}\in\mathcal{F}} \frac{1}{T} \sum_t \left(\mathbb{E} f_t(X_t) - \frac{1}{n} \sum_i f_t(X_{ti}) \right)$$

Multitask dictionary learning



$$\begin{aligned}
\mathcal{D} &= \text{dictionaries} = \{x \in H \mapsto (\langle d_1, x\rangle, ..., \langle d_K, x\rangle) : d_k \in H, \|d_k\| \leq 1\} \\
\mathcal{W} &= \text{matrices} = \left\{W \in \mathbb{R}^{TK} : \|W\|_{\mathcal{W}} \leq 1\right\} \ (\|.\|_{\mathcal{W}} \text{ to be specified}) \\
\mathcal{F} &= \mathcal{W} \circ \mathcal{D}
\end{aligned}$$

A bound for multitask dictionary learning

If $|ext(\mathcal{W})| < \infty$ we can use our trick

$$\begin{aligned}
\mathcal{R}(\mathcal{W} \circ \mathcal{D}, \mathbf{x}) &= \mathcal{R}(ext(\mathcal{W}) \circ \mathcal{D}, \mathbf{x}) \\
&= \mathcal{R}\left(\bigcup_{W \in ext(\mathcal{W})} W \circ \mathcal{D}, \mathbf{x}\right) \\
&\leq \mathfrak{S} + 8\,\mathfrak{W}\sqrt{\frac{\ln |ext(\mathcal{W})|}{nT}}
\end{aligned}$$

Typically, as $T \to \infty$, $\mathfrak{S} \to 0$ and $\mathfrak{W}$ is related to $\lambda_{\max}$ of some covariance.

Example: the sharing norm

$$
\begin{aligned}
\|W\|_{\mathcal{W}} &= \sum_k \max_t |W_{tk}| \,, \\
|ext\left(\mathcal{W}\right)| &= 2^T K \\
\mathcal{R}\left(\mathcal{W} \circ \mathcal{D}, \mathbf{x}\right) &\leq \mathfrak{S} + 8\mathfrak{W}\sqrt{\frac{\ln 2}{n} + \frac{\ln K}{nT}} \\
\mathfrak{S} &\leq \sqrt{\frac{tr\left(\hat{C}\left(\mathbf{x}\right)\right)}{nT}} \text{ and } \mathfrak{W} = \sqrt{\lambda_{\max}\left(\hat{C}\left(\mathbf{x}\right)\right)}
\end{aligned}
$$

Very weak dependence on K. Disappears as $T \to \infty$.

Other applications

Independent sparse weights: $\|W\|_{\mathcal{W}} = \max_t \sum_k |W_{tk}|$

Subspace learning: $\|W\|_{\mathcal{W}} = \max_t \sqrt{\sum_k W_{tk}^2}$ and dictionary orthogonal

To explore

$\|w\|_p$ instead of $\|w\|_2$-constraint on $\mathcal{F}_m$

Find applications with countable number of $\mathcal{F}_m$

Thank you!