Analysis of the oracle risk in multi-task kernel ridge regression

or

Does multi-task work?

Matthieu Solnon

Advisors: Sylvain Arlot & Francis Bach
Sierra team
Département d’Informatique de l’École Normale Supérieure
(CNRS/ENS/INRIA UMR 8548)

StatMathAppli 2013, Fréjus

http://www.di.ens.fr/~solnon/
matthieu.solnon@ens.fr
Multi-task dream
Multi-task dream
Set up

Design: \( n \) points \((X_i)_{i=1}^n \in \mathcal{X}^n\).

Target functions: \( p \) functions \((F^j)_{j=1}^p \in \mathcal{F}^p\).

Fixed design: the goal is to estimate \((F^j(X_i))_{i,j} = (f^j_i)_{i,j}\), we treat \((X_i)_{i=1}^n\) as fixed.

Observations: \( Y_i^j = F^j(X_i) + \varepsilon_i^j, \forall i (\varepsilon_i^j)_{j=1}^p \sim \mathcal{N}(0, \sigma^2 I_p)\).

Quadratic risk: ideally minimize, over \((G^j)_{j=1}^p \in \mathcal{F}^p\),

\[
\mathbb{E} \left[ \frac{1}{np} \sum_{i,j} (F^j(X_i) - G^j(X_i))^2 \right] = \mathbb{E} \left[ \frac{1}{np} \sum_{i,j} (f^j_i - g^j_i)^2 \right] = \mathbb{E} \left[ \frac{1}{np} \| f - g \|^2 \right]
\]
Extension of kernel ridge regression to a multi-task setting

$\mathcal{F}$ has a RKHS structure, with scalar product $\langle \cdot, \cdot \rangle_\mathcal{F}$, norm $\| \cdot \|_\mathcal{F}$, kernel $k$ and $n \times n$ kernel matrix $K$: $K_{i,j} = k(X_i, X_j)$.

**Single-task regularization**: For every task $j \in \{1, \ldots, p\}$,

regularize the empirical risk by $\lambda^j \| G^j \|_\mathcal{F}^2$.

**Multi-task regularization**: Following Evgeniou et al. (2005),

regularize the empirical risk by $\sum_{j,k} M_{j,k} \langle G^j, G^k \rangle_\mathcal{F}$.

$p \times p$ matricial regularization parameter (to calibrate): $M$.
Extension of kernel ridge regression to a multi-task setting

\( \mathcal{F} \) has a RKHS structure, with scalar product \( \langle \cdot, \cdot \rangle_{\mathcal{F}} \), norm \( \| \cdot \|_{\mathcal{F}} \), kernel \( k \) and \( n \times n \) kernel matrix \( K \): \( K_{i,j} = k(X_i, X_j) \).

**Single-task regularization**: For every task \( j \in \{1, \ldots, p\} \),
regularize the empirical risk by \( \lambda^j \| G^j \|_{\mathcal{F}}^2 \).

**Multi-task regularization**: Following Evgeniou et al. (2005),
regularize the empirical risk by

\[
\sum_{j,k} M_{j,k} \langle G^j, G^k \rangle_{\mathcal{F}}.
\]
Extension of kernel ridge regression to a multi-task setting

$\mathcal{F}$ has a RKHS structure, with scalar product $\langle \cdot, \cdot \rangle_\mathcal{F}$, norm $\| \cdot \|_\mathcal{F}$, kernel $k$ and $n \times n$ kernel matrix $K$: $K_{i,j} = k(X_i, X_j)$.

**Single-task regularization**: For every task $j \in \{1, \ldots, p\}$, regularize the empirical risk by $\lambda^j \|G^j\|^2_\mathcal{F}$.

**Multi-task regularization**: Following Evgeniou et al. (2005), regularize the empirical risk by

$$\sum_{j,k} M_{j,k} \langle G^j, G^k \rangle_\mathcal{F}.$$ 

$p \times p$ matricial regularization parameter (to calibrate): $M$
Example 1: single-task

To recover the single-task estimator, take

$$M_{\text{ind}}(\lambda^1, \ldots, \lambda^p) = \text{Diag}(\lambda^1, \ldots, \lambda^p).$$

The estimator has to be calibrated over

$$\mathcal{M}_{\text{ind}} = \left\{ M_{\text{ind}}(\lambda^1, \ldots, \lambda^p), \ (\lambda^1, \ldots, \lambda^p) \in \mathbb{R}^+_p \right\}.$$
Example 2: 1 cluster

Suppose the $p$ regression functions $F^j$ are all “close” in $\mathcal{F}$. It seems natural to regularize the empirical risk by

$$\lambda \sum_{j=1}^{p} \left\| G^j \right\|_{\mathcal{F}}^2 + \frac{\mu}{2} \sum_{j,k} \left\| G^j - G^k \right\|_{\mathcal{F}}^2.$$

This leads to $M_{SD}(\lambda, \mu) = (\lambda + p\mu)I_p - \mu 11^\top$. 
Example 2: 1 cluster

Suppose the $p$ regression functions $F^j$ are all “close” in $\mathcal{F}$. It seems natural to regularize the empirical risk by

$$
\lambda \sum_{j=1}^{p} \| G^j \|_F^2 + \frac{\mu}{2} \sum_{j,k} \| G^j - G^k \|_F^2.
$$

This leads to $M_{SD}(\lambda, \mu) = (\lambda + p\mu)I_p - \mu \mathbf{1}\mathbf{1}^\top$.

It is actually better to use

$$
M_{AV}(\lambda, \mu) = \frac{\lambda \mathbf{1}\mathbf{1}^\top}{p} + \frac{\mu}{p} \left( I_p - \frac{\mathbf{1}\mathbf{1}^\top}{p} \right).
$$

The estimator has to be calibrated over

$$
M_{AV} = \{ M_{AV}(\lambda, \mu), (\lambda, \mu) \in \mathbb{R}^2_+ \}.
$$
Example 2: 1 cluster

Suppose the $p$ regression functions $F^j$ are all “close” in $\mathcal{F}$. It seems natural to regularize the empirical risk by

$$
\lambda \sum_{j=1}^{p} \|G^j\|_{\mathcal{F}}^2 + \frac{\mu}{2} \sum_{j,k} \|G^j - G^k\|_{\mathcal{F}}^2.
$$

This leads to $M_{SD}(\lambda, \mu) = (\lambda + p\mu)l_p - \mu 11^\top$.

It is actually better to use

$$
M_{AV}(\lambda, \mu) = \frac{\lambda}{p} 11^\top + \frac{\mu}{p} \left( l_p - \frac{11^\top}{p} \right) \quad \text{(Mean)} + \frac{\mu}{p} \left( l_p - \frac{11^\top}{p} \right) \quad \text{(Variance)}.
$$

The estimator has to be calibrated over

$$
\mathcal{M}_{AV} = \{ M_{AV}(\lambda, \mu), \ (\lambda, \mu) \in \mathbb{R}_+^2 \}.
$$
Example 3: several clusters

Suppose the $p$ regression functions $F^j$ are split into several clusters.

- If the allocation of each function to the cluster is known, it is easy to adapt the previous example to this case.
- Otherwise, it becomes problematic to calibrate the matricial regularization parameter.
This regularization scheme is:

- useful to express an Hilbertian assumption on the tasks (the regression functions are clustered);
- not suited to express assumptions like sparsity or low rank.
Calibration

Previous work (Solnon et al., 2012):

- Calibration over sets like $\mathcal{M}_{\text{ind}}$ or $\mathcal{M}_{\text{AV}}$ is possible;
- The estimator verifies an oracle inequality, that is,

$$
\frac{1}{np} \mathbb{E} \left[ \left\| \hat{f}_M - f \right\|^2 \right]
$$

is close to

$$
\frac{1}{np} \inf_{M \in \mathcal{M}} \left\{ \mathbb{E} \left[ \left\| \hat{f}_M - f \right\|^2 \right] \right\}.
$$

Oracle risk
Does it really work?

We want to answer the following questions.

- Can we compare the multi-task and single task oracle risks using those parameter sets?
- Are there settings where the multi-task oracle is intrinsically better than the single-task one, or *vice versa*?
We use $\mathcal{M}_{AV}$ for the multi-task regularization, that is, we suppose that the $p$ regression functions are grouped in one cluster.

Bias-variance decomposition: the elements of interest are

- the eigenvalues of the kernel matrix $K : (\gamma_i)_{i=1}^n$;
- the coefficients of the mean of the input signals $f^j$ on the same basis that diagonalized $K : \left( \frac{\mu_i}{p} \right)_{i=1}^n$;
- the coefficients of the “variance” of the input signals $f^j$ on the same basis that diagonalized $K : (\varsigma_i^2)_{i=1}^n$. 

Assumptions

Regularity of the kernel:

\[ \forall i \in \{1, \ldots, n\}, \quad \gamma_i = ni^{-2\beta}. \quad (1) \]

Regularity of the signal:

\[ \forall i \in \{1, \ldots, n\}, \quad \begin{cases} 
\frac{\mu_i^2}{\mu_i^2} = C_1 ni^{-2\delta} \\
\frac{\mu_i^2}{\nu_i^2} = C_2 ni^{-2\delta} 
\end{cases} \quad . \quad (2) \]
Control of the multi-task oracle risk

Denote by $\mathcal{R}_{MT}^*$ the multi-task oracle risk, and by $\kappa(\beta, \delta)$ a constant depending only on $\beta$ and $\delta$.

**Theorem**

For every $n$, $p$, $C_1$, $C_2$, $\sigma^2$, $\beta$ and $\delta$ such that Eq. (1) and Eq. (2) hold and such that $1 < 2\delta < 4\beta + 1$, we have

$$
\mathcal{R}_{MT}^* \leq 2^{1/(2\delta)} \left( \frac{np}{\sigma^2} \right)^{1/(2\delta)-1} \kappa(\beta, \delta) \left[ C_1^{1/(2\delta)} + (p - 1)^{1-(1/2\delta)} C_2^{1/2\delta} \right].
$$

Moreover, there exists constants $N$ and $\alpha \in (0, 1)$ such that, if $n \geq N$, $p/\sigma^2 \leq n$ and $2 < 2\delta < 4\beta$, we have

$$
\mathcal{R}_{MT}^* \geq \alpha \left( \frac{np}{\sigma^2} \right)^{1/(2\delta)-1} \kappa(\beta, \delta) \left[ C_1^{1/2\delta} + (p - 1)^{1-(1/2\delta)} C_2^{1/2\delta} \right].
$$
Single-task assumptions

2 points: suppose for simplicity that $p$ is even and that

$$f^1 = \ldots = f^{p/2} \quad \text{and} \quad f^{p/2+1} = \ldots = f^p .$$

(2Points)

1 Outlier:

$$f^1 = \ldots = f^{p-1} .$$

(1Out)
So?

When the tasks are very similar: that is, most of the signal is in the mean, in both cases, multi-task works as if we have $p$ times more observations.

When the tasks are very different: that is, most of the signal is in the variance,
- if Eq. (2Points) holds, multi-task and single task work similarly;
- if Eq. (1Out) holds, multi-task does arbitrarily worse than single-task.
Further relaxation of the assumptions to get one group of tasks.

Figure: Setting C.
Relaxation of (1Out)

Figure: Relaxation of Assumption (1Out) (Setting D).
Conclusion

- It can work.
- The procedure can be affected by (even slight) modelisation errors.

What next?

- Design a more robust procedure (for instance, able to choose between $\mathcal{M}_1$ and $\mathcal{M}_2$ when $\mathcal{M}_1 \subset \mathcal{M}_2$).
- For a given distribution of the $p$ regression functions, what would be the oracle matrix $\mathcal{M}^* \in S_p^+ (\mathbb{R})$? Can we find, at least, a set of candidate matrices $\mathcal{M}^* \subset S_p^+ (\mathbb{R})$?
- We might need new concentration results and optimization tools.
The end

Thanks!

