An Efficient Method for Statistical Learning by Means of Tensor Format Representations

Mike Espig

RWTH Aachen University
Department of Mathematics, IGPM
Numerical and Applied Analysis

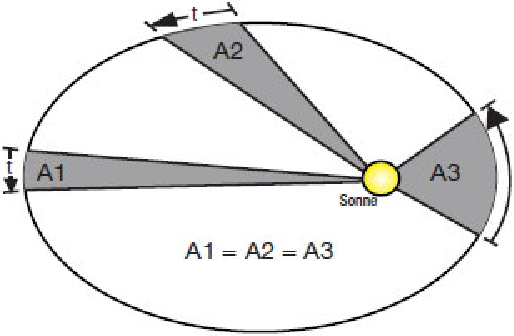
Poperschau
20.09.2016

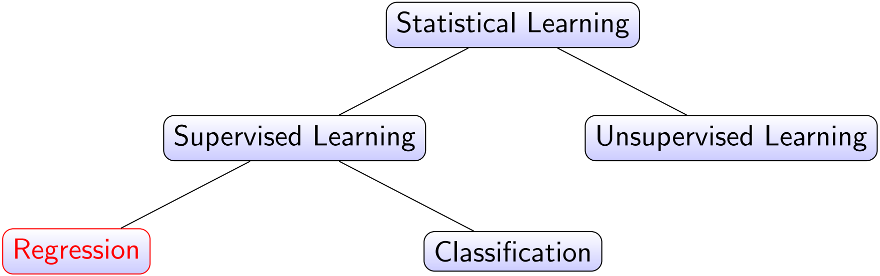
Outline

1. Introduction
2. Tensor Formats
3. Alternating Steepest Descent Algorithm (ASD)
4. Convergence of the ASD & PASD Method
5. Numerical Experiment

Introduction

The problem of searching for patterns in data is a fundamental one and has a long and successful history. For example, the extensive astronomical observations of Tycho Brahe in the 16th century allowed Johannes Kepler to discover the empirical laws of planetary motion, which in turn provided a springboard for the development of classical mechanics.




Statistical Learning
Supervised Learning
Unsupervised Learning
Regression
Classification

Regression (Problem Setting)

Sample set of independent and identically distributed (i.i.d.) observations drawn according to $\mathcal{P}(x,y) = \mathcal{P}(x)\mathcal{P}(y|x)$,

$$\mathcal{S} = \Big\{(x_\ell, y_\ell) \in \mathcal{X}^d \times \mathcal{Y} \mid 1 \leq \ell \leq m\Big\}$$

Set of hypotheses

$$\mathcal{H} = \{h(\cdot, p) : \mathcal{X}^d \to \mathcal{Y} \mid p \in P\}$$

Risk functional

$$R(p) = \int_{\mathcal{Z}^d} (h(x_\ell, p) - y_\ell)^2 \mathrm{d}\mathcal{P}(x,y)$$

The regression function is the one that minimises the risk functional,

$$\varrho(x) = \int_{\mathcal{Y}} y \, \mathrm{d}\mathcal{P}(y|x).$$

But in our situation, the joint probability distribution function $\mathcal{P}(x,y)$ is unknown.

Regression Estimation

Remark

If the regression function ϱ does not belong to the set of hypotheses $\mathcal{H}$, then the function $h(\cdot, p^*) \in \mathcal{H}$ minimising the risk functional is the closest to the regression in the metric

$$\nu(\varrho, h(\cdot, p^*)) = \sqrt{\int_{\mathcal{X}^d} (\varrho(x) - h(x, p^*))^2 \, \mathrm{d}\mathcal{P}(x)},$$

where the existence of $h(\cdot, p^*)$ is provided.

Empirical Risk Minimization Principle

In order to minimise the risk functional R with an unknown distribution function $\mathcal{P}(x, y)$, the following inductive principle is applied in statistical learning:

(i) The risk functional R is replaced by the so-called empirical risk functional

$$R_{\text{emp}}(p) = \frac{1}{m} \sum_{(x,y)\in\mathcal{S}} (h(x,p) - y)^2$$

constructed on the basis of the finite sample set $\mathcal{S}$.

(ii) One approximates the hypothesis $h(\cdot, p^*)$ that minimise the risk functional R by the function $h(\cdot, p_{\mathcal{S}}) \in \mathcal{H}$ minimising the empirical risk.

This principle is called the empirical risk minimisation (ERM) principle.

Performance of the Estimator

Theorem (See e.g. Györfi et al. (2002))

Let $1 \leqslant L < \infty$*. Assume* $|y| \leqslant L$ *almost surely. Let the estimate* h_{emp} *be defined by minimization of the empirical risk over a set of functions* $\mathcal{H}$ *and truncation at* $\pm L$*. Then one has*

$$E\{\int_{\mathcal{X}^d}(h_{\mathrm{emp}}(x,p)-\varrho(x))^2 \mathrm{d}\mathcal{P}(x)\}$$

$$\leqslant \frac{c_1}{m} + \frac{(c_2 + c_3\log(m))V_{\mathcal{H}^+}}{m} + 2\inf_{h\in\mathcal{H}}\int_{\mathcal{X}^d}(h(x,p)-\varrho(x))^2\mathrm{d}\mathcal{P}(x),$$

where

$$c_1 = 24{\cdot}214L^4(1{+}\log 42),\ c_2 = 48{\cdot}214L^4\log(480eL^2),\ c_3 = 48{\cdot}214L^4$$

and $V_{\mathcal{H}^+}$ *is the* VC *dimension of* $\mathcal{H}_+ := \{\mathrm{hyp}(h) \mid h \in \mathcal{H}\}$
(Vapnik - Chervonenkis).

Contributions

Every hypothesis $h(\cdot, p)$ is a linear combination of elementary features, i.e.

$$h(x,p) = \sum_{i_1=1}^{n_1} \cdots \sum_{i_d=1}^{n_d} c(p)_{(i_1,\dots,i_d)} \prod_{\nu=1}^{d} \varphi_{\nu,i_\nu}(x_\nu) = \left\langle c(p), \bigotimes_{\nu=1}^{d} \varphi_\nu(x_\nu) \right\rangle,$$

where $c(p) \in \mathbb{R}^{n_1 \times \cdots \times n_d}$ and
$\varphi_\nu(x_\nu) = \left(\varphi_{\nu,1}(x_\nu), \cdots, \varphi_{\nu,n_\nu}(x_\nu)\right)^T \in \mathbb{R}^{n_\nu}$.
In our new ansatz, the coefficient tensor $c(p)$ is represented in a tensor format.

$$\Rightarrow R_{\text{emp}}(p) = \frac{1}{2} \langle Ac(p), c(p) \rangle - \langle b, c(p) \rangle + const.$$

$A \geq 0,\ A = A^t$.

Tensor Formats

What is a Tensor Format Representation?

$$U : \bigtimes_{\mu=1}^{L} P_\mu \to \bigotimes_{\nu=1}^{d} \mathbb{R}^{n_\nu} \qquad (L \geq d)$$

$u = U(p_1, \dots, p_L)$ is represented in the tensor format U

- U is multilinear in $p_1, \dots, p_L$
- $(p_1, \dots, p_L)$ is a *representation system* of u

Example (r-Term Representation)

$p_\mu = (p_{\mu,j} \in \mathbb{R}^n \,:\, 1 \leq j \leq r)$

$$(p_1, \dots, p_d) \mapsto U_{r\text{-term}}(p_1, \dots, p_d) = \sum_{j=1}^{r} \bigotimes_{\mu=1}^{d} p_{\mu,j}$$

Example (Matrix Product States (MPS), Tensor-Train (TT))

$$u_{\underline{p}} = \sum_{j_1=1}^{r} \sum_{j_2=1}^{r} \sum_{j_3=1}^{r} p_{1,j_1} \otimes p_{2,j_1,j_2} \otimes p_{3,j_2,j_3} \otimes p_{4,j_3} \quad (p_{\mu,\cdot} \in \mathbb{R}^n)$$

Example (Conformal Tensor Formats)

Quantum Mechanics (two-electron integrals):

$$u_{(w,\underline{p})} = \sum_{j_1=1}^{r} \sum_{j_2=1}^{r} w_{j_1,j_2} \cdot p_{1,j_1} \otimes p_{2,j_2} \otimes p_{3,j_1} \otimes p_{4,j_2}, \; (p_{\mu,\cdot} \in \mathbb{R}^n, \, w_{\cdot} \in \mathbb{R})$$

Alternating Steepest Descent Algorithm

Optimisation Respect to Parameter Space

$$f(u) = \frac{1}{2}\langle Au, u\rangle - \langle b, u\rangle,$$

- u is substituted by a tensor representation: $u := U(p_1, \ldots, p_L)$

$$\Rightarrow f(u) = F(p_1, \ldots, p_L) = f(U(p_1, \ldots, p_L))$$

- We are looking for a representation system $(p_1^*, \ldots, p_L^*)$ such that

$$F(p_1^*, \ldots, p_L^*) = \inf_{(p_1, \ldots, p_L) \in P} F(p_1, \ldots, p_L).$$

Alternating Steepest Descent Algorithm

$$\begin{aligned} v &= U(p_1, \ldots, , p_{\mu-1}, p_\mu, p_{\mu+1}, \ldots, p_L) \\ &= W_\mu(p_1, \cdots, p_{\mu-1}, p_{\mu+1}, \ldots, p_L) p_\mu \\ &=: W_\mu p_\mu \end{aligned}$$

The following holds:

(i) W_μ is a linear map and $\mathrm{ran}\,(W_\mu)$ is a linear subspace of $\bigotimes_{\mu=1}^{d} \mathbb{R}^n$.

(ii) $W_\mu \subset \mathrm{ran}\,(U)$, i.e. addition of represented tensors in W_μ will not change the ranks

(ii) Direction of steepest ascent in $U_\mu := \mathrm{span}(W_\mu)$

$$\frac{1}{\|W_\mu^T(Av-b)\|_{P_\mu}} W_\mu W_\mu^T(Av-b) = \mathrm{argmax}_{W_\mu q_\mu \in U_\mu} \frac{\langle f'(v), W_\mu q_\mu \rangle}{\|q_\mu\|_{P_\mu}}$$

Alternating Steepest Descent Algorithm

Algorithmus 1 ASD method

Choose initial $p^1 \in P$, and define $k := 1$.
while Stop Condition **do**
    for $1 \leq \mu \leq L$ **do**

$$
\begin{aligned}
r_{k,\mu} &:= b - Av_{k,\mu} \\
d_{k,\mu} &:= W_{k,\mu} M_{k,\mu}^{-1} W_{k,\mu}^T r_{k,\mu} \\
\lambda_{k,\mu} &:= \frac{\langle r_{k,\mu}, d_{k,\mu} \rangle}{\langle A d_{k,\mu}, d_{k,\mu} \rangle} \\
v_{k,\mu+1} &:= v_{k,\mu} + \lambda_{k,\mu} d_{k,\mu}
\end{aligned}
$$

    end for
    $k \mapsto k + 1$.
end while

Pivotised Alternating Steepest Descent Algorithm

Algorithmus 2 Pivotised ASD (PASD) method

Choose initial $p^1 \in P$, and define $k := 1$.

while Stop Condition **do**

  $\mu := \operatorname{argmax}_{1 \leq \nu \leq L} \left\| \frac{\partial F}{\partial p_\nu}(p_{k,\nu}) \right\|_\infty$

$$
\begin{aligned}
r_{k,\mu} &:= b - A v_{k,\mu} \\
d_{k,\mu} &:= W_{k,\mu} M_{k,\mu}^{-1} W_{k,\mu}^T r_{k,\mu} \\
\lambda_{k,\mu} &:= \frac{\langle r_{k,\mu}, d_{k,\mu} \rangle}{\langle A d_{k,\mu}, d_{k,\mu} \rangle} \\
v_{k,\mu+1} &:= v_{k,\mu} + \lambda_{k,\mu} d_{k,\mu}
\end{aligned}
$$

  $k \mapsto k + 1$.

end while

The Alternating Least Squares Algorithm

Algorithmus 3 ALS method

1: Choose initial $p^1 \in P$, and define $k := 1$.
2: **while** Stop Condition **do**
3:     **for** $1 \leq \mu \leq L$ **do**
4:         Compute the minimum norm solution of the least squares problem

$$p_\mu^{k+1} := \operatorname{argmin}_{q_\mu \in P_\mu} F(p_1^{k+1}, \ldots, p_{\mu-1}^{k+1}, q_\mu, p_{\mu+1}^k, \ldots, p_L^k).$$

5:     **end for**
6:     $k \mapsto k + 1$.
7: **end while**

Numerical Cost of ALS = Numerical Cost of ASD + $\mathcal{O}(m^3)$

$m := \max_{1 \leq \mu \leq L} \dim P_\mu$

Convergence of the ASD & PASD Method

Notation

$\left(u^k\right)_{k\in\mathbb{N}} \subset \mathcal{V}$ is the sequence of corresponding tensors from the ALS iteration, i.e.

$$u^k := U(p^k) \quad \text{for all } k \in \mathbb{N}.$$

The set of accumulation points of $\left(u^k\right)_{k\in\mathbb{N}}$ is denoted by $\mathcal{A}\left(u^k\right)$.

Critical Points

The set of *critical points* $\mathfrak{M}$ is defined by

$$\mathfrak{M} := \{u \in \mathcal{V} \mid \exists p \in P : u = U(p) \wedge F'(p) = 0\}.$$

Convergence of the ASD & PASD Method

General Assumption

- Suppose that the sequence of parameter $\left(p^k\right)_{k\in\mathbb{N}} \subset P$ is bounded.
- For all $\mu \leq L$ there exists k_0 and $\gamma_\mu > 0$ such that for all $k \geq k_0$ we have

$$\sigma^{[\mu]}_{\min,+}(W_{k,\mu}) := \min\left\{\sigma_{k,\mu} > 0 : \sigma_{k,\mu} \text{ is singular value of } W_{k,\mu}\right\} \geq \gamma_\mu.$$

The assumptions are motivated by the counterexample of Lim and de Silva (2008).

$$b = x \otimes x \otimes y + x \otimes y \otimes x + y \otimes x \otimes x$$

$$v_k = \left(x + \frac{1}{k}\,y\right) \otimes \left(x + \frac{1}{k}\,y\right) \otimes (kx + y) - x \otimes x \otimes kx \xrightarrow[k\to\infty]{} b.$$

Convergence of the ASD & PASD Method

$$\tan^2 \angle[\bar{u}, u_{k,\mu+1}] = \left(\frac{q_{k,\mu}^{(s)}}{q_{k,\mu}^{(c)}} \right)^2 \tan^2 \angle[\bar{u}, u_{k,\mu}],$$

where

$$d_{k,\mu} = \left[\begin{array}{cc} \bar{u} & R \end{array} \right] \left(\begin{array}{c} \gamma_{k,\mu} \\ \rho_{k,\mu} \end{array} \right), \quad u_{k,\mu} = \left[\begin{array}{cc} \bar{u} & R \end{array} \right] \left(\begin{array}{c} c_{k,\mu} \\ s_{k,\mu} \end{array} \right)$$

$$\begin{aligned} q_{k,\mu}^{(s)} \quad &:= \quad \frac{\|s_{k,\mu} + \lambda_{k,\mu}\rho_{k,\mu}\|}{\|s_{k,\mu}\|} \\ q_{k,\mu}^{(c)} \quad &:= \quad \frac{|c_{k,\mu} + \lambda_{k,\mu}\gamma_{k,\mu}|}{|c_{k,\mu}|} \end{aligned}$$

Convergence of the ASD & PASD Method

RWTH AACHEN UNIVERSITY

Theorem (E., (2016))

- *Every accumulation point of* $\left(u^k\right)_{k\in\mathbb{N}} \subset \mathcal{V}$ *is a critical point, i.e.* $\mathcal{A}\left(u^k\right) \subseteq \mathfrak{M}$*, furthermore*

$$\operatorname{dist}(u^k, \mathfrak{M}) \xrightarrow[k\to\infty]{} 0.$$

- $$u^k \xrightarrow[k\to\infty]{} \bar{u} \quad \textit{for PASD}$$

  and if one accumulation point $\bar{u}$ *is isolated, then*

$$u^k \xrightarrow[k\to\infty]{} \bar{u} \quad \textit{for ASD,}$$

where

$$\tan\angle[u_{k,\mu+1}, \bar{u}] \le q_\mu \tan\angle[u_{k,\mu}, \bar{u}],$$

with $q_\mu := \limsup_{k\to\infty} \left|\frac{q^{(s)}_{k,\mu}}{q^{(c)}_{k,\mu}}\right|$.

Numerical Experiment

$$
\begin{aligned}
U(p_1, \ldots, p_d) &= p_1 \otimes \ldots \otimes p_d, \\
f(u) &= \frac{1}{2}\|u\|^2 + \langle b, u \rangle \quad (A = \mathsf{id}) \\
\\
\langle p, q \rangle &= 0,\ \|p\| = 1,\ \|q\| = 1 \\
b_\lambda &= \bigotimes_{\mu=1}^{3} p + \lambda \left(p \otimes q \otimes q + q \otimes p \otimes q + q \otimes q \otimes p \right),
\end{aligned}
$$

Numerical Experiment

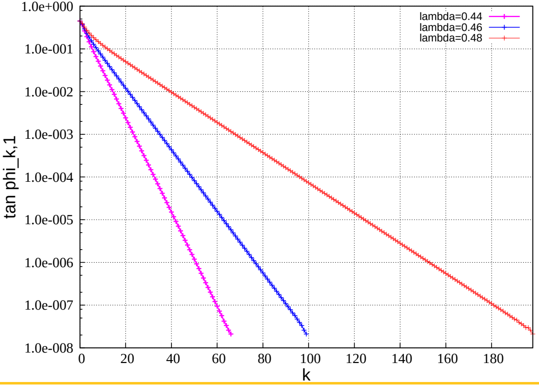
$$\begin{aligned} U(p_1, \ldots, p_d) &= p_1 \otimes \ldots \otimes p_d, \\ f(u) &= \frac{1}{2}\|u\|^2 + \langle b, u\rangle \quad (A = \mathsf{id}) \\ \\ \langle p, q\rangle &= 0,\ \|p\| = 1,\ \|q\| = 1 \\ b_\lambda &= \bigotimes_{\mu=1}^{3} p + \lambda\left(p \otimes q \otimes q + q \otimes p \otimes q + q \otimes q \otimes p\right), \\ \limsup_{k\to\infty} \left|\frac{q_{k,\mu}^{(s)}}{q_{k,\mu}^{(c)}}\right| &= \frac{\lambda}{2}\left(3\lambda + \lambda^2 + \sqrt{(3\lambda + \lambda^2)^2 + 4\lambda}\right) \end{aligned}$$

Note: The ALS method has the same rate of convergence, E. Khachatryan (2014).

Numerical Experiment

$\lambda < \frac{1}{2}$

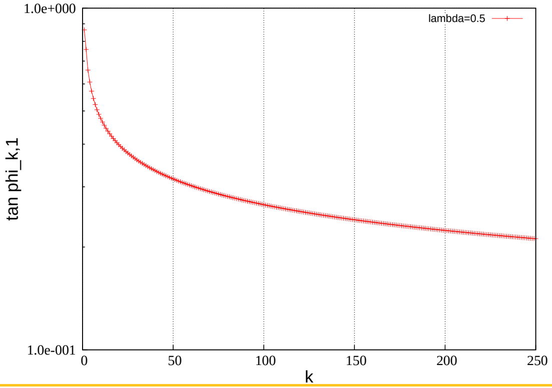
$$\limsup_{k\to\infty} \left| \frac{q_{k,\mu}^{(s)}}{q_{k,\mu}^{(c)}} \right| < 1$$



Numerical Experiment

$\lambda = \frac{1}{2}$

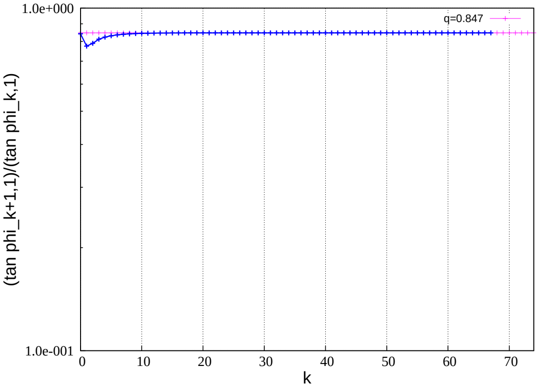
$$\limsup_{k\to\infty} \left| \frac{q_{k,\mu}^{(s)}}{q_{k,\mu}^{(c)}} \right| = 1$$



Numerical Experiment

$\lambda = 0.46$

$$\limsup_{k\to\infty} \left| \frac{q^{(s)}_{k,\mu}}{q^{(c)}_{k,\mu}} \right| \doteq 0.847$$



Numerical Experiment

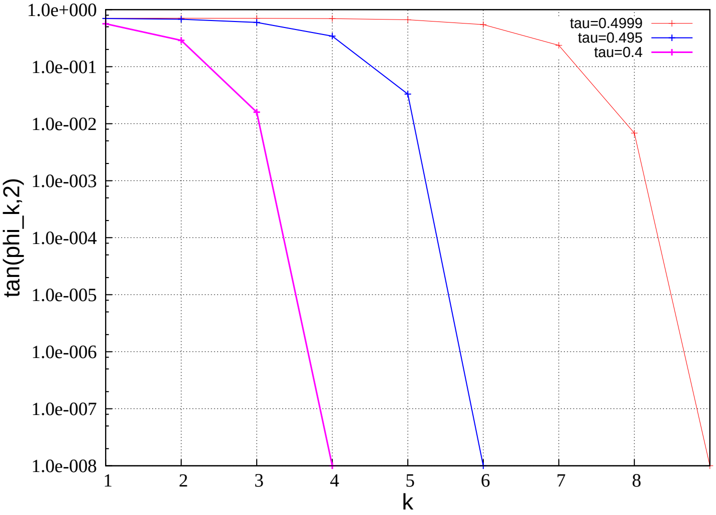
$$\begin{aligned} U(p_1, \dots, p_d) &= p_1 \otimes \dots \otimes p_d \\ b &= \sum_{j=1}^{r} \lambda_j \bigotimes_{\mu=1}^{d} b_{j\mu}, \quad \lambda_1 \geq \dots \geq \lambda_r > 0, \langle b_{i\mu}, b_{j\mu} \rangle = \delta_{ij} \end{aligned}$$

Numerical Experiment

$$
\begin{aligned}
U(p_1,\ldots,p_d) &= p_1 \otimes \ldots \otimes p_d \\
b &= \sum_{j=1}^{r} \lambda_j \bigotimes_{\mu=1}^{d} b_{j\mu}, \quad \lambda_1 \geq \cdots \geq \lambda_r > 0, \langle b_{i\mu}, b_{j\mu} \rangle = \delta_{ij}
\end{aligned}
$$

$$
q_\mu = \limsup_{k\to\infty} \left| \frac{q_{k,\mu}^{(s)}}{q_{k,\mu}^{(c)}} \right| = 0
$$

Numerical Experiment



Numerical Experiment: Statistical Learning

(Joint work with L. Sobolevskaya)

Error measure

$$\mathrm{RMSD} = \sqrt{\frac{\sum_{\ell=1}^{m}(h(x,p^*) - y_\ell)^2}{m}}.$$

Yacht Problem

This problem consists of predicting the residuary resistance of sailing yachts at the initial design stage. This data set comprises $m = 308$ full-scale experiments.

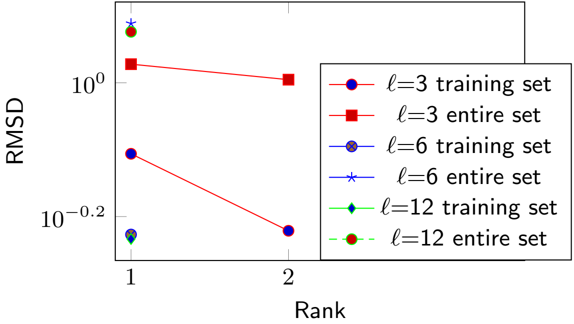
1. Longitudinal position of the center of buoyancy.
2. Prismatic coefficient.
3. Length-displacement ratio.
4. Beam-draught ratio.
5. Length-beam ratio.
6. Froude number.

The output variable is the residuary resistance per unit weight of displacement:

1. Residuary resistance per unit weight of displacement.

Numerical Experiment: Yacht Problem

Spline bases functions.



	$\ell = 3$	$\ell = 6$	$\ell = 12$
$RMSD_{TS}$	0.600926	0.593134	0.584214
$RMSD_{ES}$	1.01085	1.22633	1.19185
Max rank	2	1	1
Runtime [sec]	4.91	3.209	2.082

Comparisons

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	1.192	3.457	4.614
Runtime	2.1 secs	1.15 mins	13.6 secs

Table: The evaluations has been performed on the sismatically chosen training sets set equal to 20% of the data set.

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	1.61907	7.765	3.705
runtime	1.443 secs	1.74 hours	31.43 secs

Table: The evaluations has been performed on the randomly chosen training sets set equal to 20% of the data set.

Comparisons

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	0.736	1.268	1.33
Runtime	4.57 secs	14.37 mins	39.53 secs

Table: The evaluations has been performed on the on the systematically chosen training set equal to 40% of the data set.

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	0.975	4.0445	1.676
Runtime	1.83 secs	43.35 mins	73 secs

Table: The evaluations has been performed on the randomly chosen training sets set equal to 40% of the data set.

Comparisons

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	0.757	1.586	1.0097
Runtime	6.669 secs	1.2 hours	76.778 secs

Table: The evaluations has been performed on the systematically chosen training set equal to 60% of the data set.

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	0.826	0.933	1.075
Runtime	3.2 secs	31.876 min	89.8 secs

Table: The evaluations has been performed on the randomly chosen training sets set equal to 60% of the data set.

Comparisons

	our approach	'neuralnet'	'scikit-learn'
$RMSD_{100\%}$	0.5809	0.587	0.713
Runtime	51.3 secs	4.975 hours	107.51 secs

Table: The evaluations has been performed on randomly chosen training sets set equal to 80% of the data set.

Numerical Experiment: Two Electron Integrals

Let $\mathfrak{B} := \{\varphi_i : \mathbb{R}^3 \to \mathbb{R} : 1 \leq i \leq k\}$ be a set of so called atomic orbitals.

Two electronic Integrals are defined by

$$t_{i_1,i_2,i_3,i_4} = c \int_{\mathbb{R}^3} \int_{\mathbb{R}^3} \frac{\varphi_{i_1}(x)\varphi_{i_2}(x)\varphi_{i_3}(y)\varphi_{i_4}(y)}{\|x-y\|} dxdy \;\; f.a. \;\; i_1, ..., i_4 \in \mathbb{N}_k$$

Let $I = \mathbb{N}_k \times \mathbb{N}_k \times \mathbb{N}_k \times \mathbb{N}_k$.

We want to approximate t_{i_1,i_2,i_3,i_4} $f.a.$ $(i_1, ..., i_4) \in I$ with a tensor of the smallest possible rank.

Random Training Sets

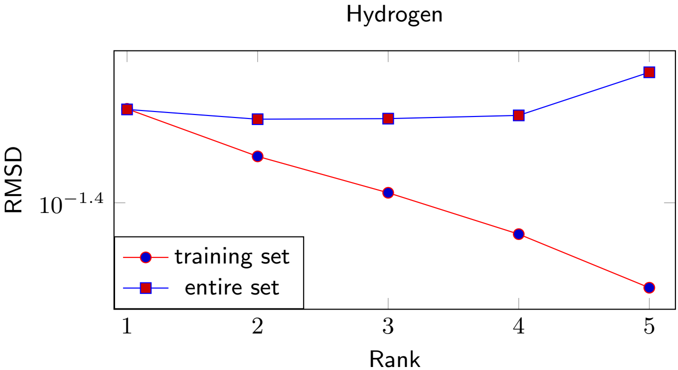


Figure: The hypotheses was trained on randomly chosen training sets equal to 50 % of the entire data set and then evaluated on the entire set. Running time was 162.972 sec

Overfitting

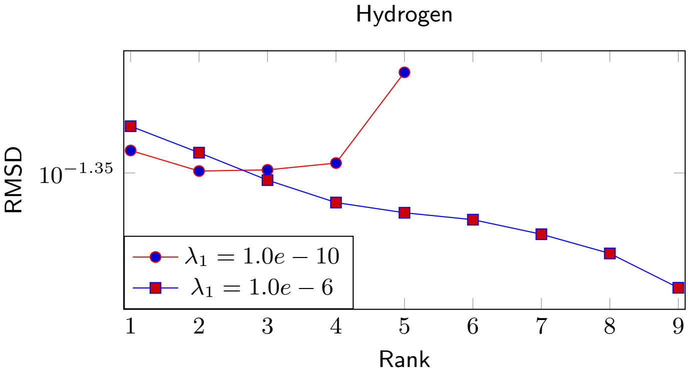


Figure: RMSD value of the entire set. As training sets were used randomly chosen 50% of the entire set. Running time was 69.1 sec

Overfitting

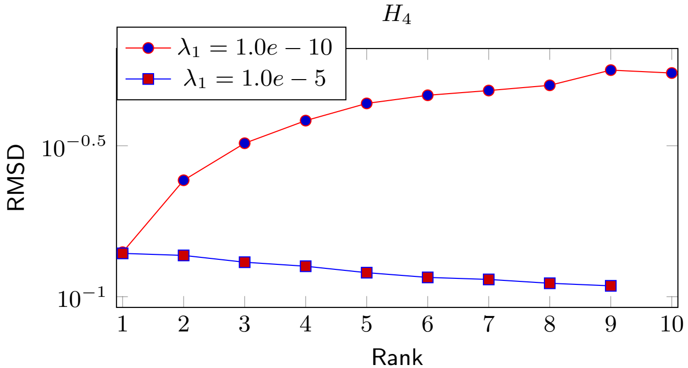
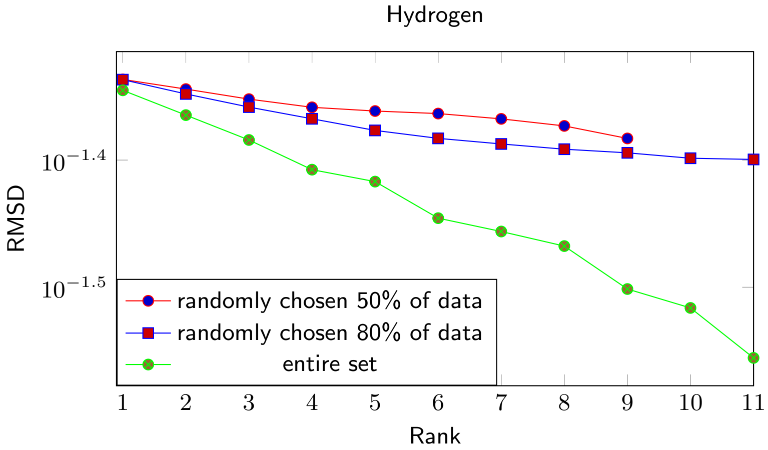


Figure: RMSD value of the entire set. As training sets were used randomly chosen 50% of the entire set.

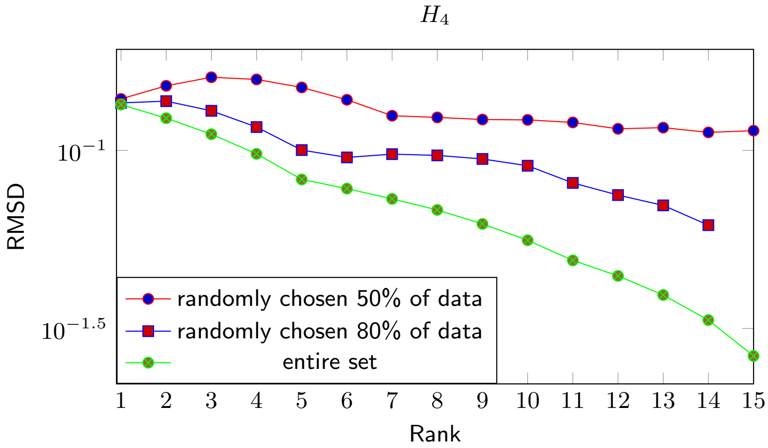
Summary

	50% of the entire set	80% of the entire set	entire set
$RMSD(I)$	0.0432166	0.0408642	0.0278161
Rank	9	11	11
Runtime [sec]	69.1	483.945	1276.86



Summary

	50% of the entire set	80% of the entire set	entire set
$RMSD(I)$	0.12461	0.0724535	0.0264707
Rank	15	14	15
Runtime [sec]	25.2576	45.9685	84.528



Publications & Source Files

- http://www.alopax.de/publications
- **Tensor Calculus**, Open Source Lib in C++, http://gitorious.org/tensorcalculus/pages/Home [H. Auer, Espig, Handschuh, Wähnert, 2011]