A Experiments on a real data set

We considered the Breast Cancer Wisconsin (Diagnostic) Data Set[7], that was considered in a similar analysis in [12]. It consists in 194 examples (4 were removed since without labels) with 32 variables. We divided the dataset into a training set of size 60, a validation set of size 60, and a test set of size 74. We experiment with polynomial tensor kernels of different degree (1,2,3,4) and we compared the test MSE for L1, L4/3, and L2 regularized least squares. See Tables 3 and 4 below. Note that the reported computing time is the total time for solving the Lp-norm regularized problem along a sequence of 100 values of the regularization parameter (and does not include the time required for computing either the kernel or the feature map, depending on the algorithm[8]).

We experimented also with nonparametric models. We indeed tested the performance of the exponential tensor kernel (L4/3) described in the paper and we compare it with the Gaussian (standard matrix) kernel (L2). We obtained a test MSE = 973.27 for the Exponential tensor kernel and a test MSE = 928.57 for the Gaussian kernel.

Table 3: (L2)RLS and LASSO(L1). Polynomial models.

degree	n. features	RLS test MSE	RLS TIME(sec)	LASSO test MSE (and selected feat.)	LASSO TIME(sec)
1	33	1240.99	< 0.1	1052.14(10)	0.4
2	561	2240.17	< 0.1	1444.84(12)	0.6
3	6545	3264.07	< 0.1	2297.36(15)	5.8
4	58905	3304.96	< 0.1	3067.24(15)	49.8

Table 4: (L4/3)RLS. Polynomial models.

degree	n. features	test MSE (feat. above 2STD)	TIME(sec) Dual Algo with $\mathbf{K}$	Dual Algo without $\mathbf{K}$	FISTA primal
1	33	1078.57(1)	9.1	0.2	0.1
2	561	2105.40(24)	7.6	0.6	0.7
3	6545	2139.64(136)	19.9	15.3	16.8
4	58905	3011.86(932)	25.5	151.8	116.3

B Proofs and technical results

This section contains proofs and additional details on some of the topics discussed above.

B.1 Duality in ℓ^p-regularization

Proof of Theorem 3.1. Problem (22) can be written in the form

$$\min_{w \in \ell^p(\mathbb{K})} f(\boldsymbol{w}) + g(-\Phi_n \boldsymbol{w}), \tag{29}$$

where $f(\boldsymbol{w}) = (1/p)\|\boldsymbol{w}\|_p^p$, $g(\boldsymbol{\beta}) = \gamma \sum_{i=1}^n L(y_i, -\beta_i)$, and Φ_n is defined as in (23). The Fenchel-Rockafellar dual problem of (29) is [1]

$$\min_{\boldsymbol{\alpha} \in \mathbb{R}^n} f^*(\Phi_n^* \boldsymbol{\alpha}) + g^*(\boldsymbol{\alpha}) \tag{30}$$

[7]Source UCI, http://archive.ics.uci.edu/ml/datasets/Breast+Cancer+Wisconsin+%28Diagnostic%29

[8]In particular, FISTA on the primal and the dual algorithm without kernel require the computation of the feature map on the training, validation, and test sets, which in the case of polynomial models of degree 4 takes about 180s.

and the corresponding KKT optimality conditions are

$$\bar{\boldsymbol{w}} \in \partial f^*(\Phi_n^*\bar{\boldsymbol{\alpha}}) \qquad \text{and} \qquad \bar{\boldsymbol{\alpha}} \in \partial g(-\Phi_n\bar{\boldsymbol{w}}).$$

Now it is easy to see that

$$(\forall\, \boldsymbol{\alpha} \in \mathbb{R}^n) \quad g^*(\boldsymbol{\alpha}) = \gamma \sum_{i=1}^n L^*\Big(y_i, -\frac{\alpha_i}{\gamma}\Big) \quad \text{and} \quad (\forall\, \boldsymbol{u} \in \ell^q(\mathbb{K})) \quad f^*(\boldsymbol{u}) = \frac{1}{q}\|\boldsymbol{u}\|_q^q.$$

Therefore, the dual form (24) follows. Statement (i) comes from the fact that g is continuous. Statement (ii) follows from the KKT conditions above by noting that f^* is indeed differentiable and $\nabla f^* = J_q$, and the fact that g is separable. □

We now specialize Theorem 3.1 to distance-based and margin-based losses [3, Definitions 2.24 and 2.32].

Corollary B.1. *Suppose that L is a convex distance-based loss of the form $L(y,t) = \psi(y-t)$ with $\mathcal{Y} = \mathbb{R}$, for some convex function $\psi\colon \mathbb{R} \to \mathbb{R}_+$. Then the dual problem* (24) *becomes*

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{q}\|\Phi_n^*\boldsymbol{\alpha}\|_q^q - \boldsymbol{y}^\top\boldsymbol{\alpha} + \gamma\sum_{i=1}^n \psi^*\Big(\frac{\alpha_i}{\gamma}\Big) \tag{31}$$

Suppose that L is a convex margin-based loss of the form $L(y,t) = \psi(yt)$ with $\mathcal{Y} = \{-1,1\}$, for some convex function $\psi\colon \mathbb{R} \to \mathbb{R}_+$. Then the dual problem (24) *becomes*

$$\min_{\alpha\in\mathbb{R}^n} \frac{1}{q}\|\Phi_n^*\boldsymbol{\alpha}\|_q^q + \gamma\sum_{i=1}^n \psi^*\Big(-\frac{y_i\alpha_i}{\gamma}\Big). \tag{32}$$

The following example shows that all the losses commonly used in machine learning admit explicit Fenchel conjugates.

Example B.2.

(i) The *least squares loss* is $L(y,t) = \psi(y-t)$ with $\psi = (1/2)|\cdot|^2$. In that case (31) reduces to (8), which is is strongly convex with modulus $1/\gamma$.

(ii) The *Vapnik-ε-insensitive loss* for regression is $L(y,t) = \psi(y-t)$ with $\psi = |\cdot|_\varepsilon$. Then, $\psi^* = \varepsilon|\cdot| + \iota_{[-1,1]}$ and the last term in (31) turns out to be $\varepsilon\|\boldsymbol{\alpha}\|_1 + \iota_{\gamma[-1,1]^n}(\boldsymbol{\alpha})$.

(iii) The *Huber loss* is the distance-based loss defined by

$$\psi(r) = \begin{cases} r^2/2 & \text{if } |r| \le \rho \\ \rho|r| - \rho^2/2 & \text{otherwise.} \end{cases}$$

Then $\psi^* = \iota_{[-\rho,\rho]} + (1/2)|\cdot|^2$ [1, Example 13.7] and the last term in (31) is $(1/(2\gamma))\|\boldsymbol{\alpha}\|_2^2 + \iota_{\rho\gamma[-1,1]^n}(\boldsymbol{\alpha})$.

(iv) The *logistic loss* for classification is the margin-based loss with $\psi(r) = \log(1+e^{-r})$. Thus

$$\psi^*(s) = \begin{cases} (1+s)\log(1+s) - s\log(-s) & \text{if } s \in\,]-1,0[\\ 0 & \text{if } s \in \{0,-1\} \\ +\infty & \text{otherwise.} \end{cases}$$

See [1, Example 13.2(vi)]. It is easy to see that ψ has Lipschitz continuous derivative with constant $1/4$ and hence ψ^* is strongly convex with modulus 4 [1]. Thus, referring to (31), we see that in this case $\operatorname{dom}\Lambda = \prod_{i=1}^n (y_i[0,\gamma])$ and Λ is differentiable on $\operatorname{int}(\operatorname{dom}\Lambda)$ with locally Lipschitz continuous gradient. Moreover, since $\lim_{s\to 1}|(\psi^*)'(s)| = \lim_{s\to 0}|(\psi^*)'(s)| = +\infty$, we have that $\|\nabla\Lambda((\alpha))\| = +\infty$ on the boundary of $\operatorname{dom}\Lambda$. Finally, it follows from (25) that $0 < y_i\bar{\alpha}_i < \gamma$, for $i = 1,\dots,n$. Note that in this case we can still apply Algorithm 3.3 with $\varphi_2 = 0$ (see [13, Section 4]).

(v) The *hinge loss* is the margin-based loss with $\psi(r) = (1-r)_+$. We have $\psi^*(s) = s + \iota_{[-1,0]}(s)$. So the second term in (32) is $-\sum_{i=1}^n y_i\alpha_i + \iota_{\gamma[0,1]}(y_i\alpha_i)$

We also note that in all cases, for every $i \in \{1,\dots,n\}, \inf L^*(y_i,\cdot) > -\infty$, which was a condition considered in Proposition 3.2.

Proof of Proposition 3.2. We use the same notation as in the proof of Theorem 3.1. It follows from the definitions of $\boldsymbol{w}$ and $\bar{\boldsymbol{w}}$ and the Young-Fenchel equalities [1] that $f(\bar{\boldsymbol{w}}) + f^*(\Phi_n^*\bar{\boldsymbol{\alpha}}) = \langle \bar{\boldsymbol{w}}, \Phi_n^*\bar{\boldsymbol{\alpha}}\rangle$ and $f(\boldsymbol{w}) + f^*(\Phi_n^*\boldsymbol{\alpha}) = \langle \boldsymbol{w}, \Phi^*\boldsymbol{\alpha}\rangle$, and hence

$$f^*(\Phi_n^*\boldsymbol{\alpha}) - f^*(\Phi_n^*\bar{\boldsymbol{\alpha}}) = f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}) + \langle \Phi_n^*\alpha, \boldsymbol{w}\rangle - \langle \Phi_n^*\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{w}}\rangle.$$

Since $-\Phi\bar{\boldsymbol{w}} \in \partial g^*(\bar{\boldsymbol{\alpha}})$, we have

$$g^*(\boldsymbol{\alpha}) - g^*(\bar{\boldsymbol{\alpha}}) \geq \langle -\Phi_n\bar{\boldsymbol{w}}, \boldsymbol{\alpha} - \bar{\boldsymbol{\alpha}}\rangle = \langle \Phi_n^*\bar{\boldsymbol{\alpha}}, \bar{\boldsymbol{w}}\rangle - \langle \Phi_n^*\boldsymbol{\alpha}, \bar{\boldsymbol{w}}\rangle.$$

Summing the two inequalities above, we get

$$\begin{aligned}\Lambda(\boldsymbol{\alpha}) - \Lambda(\bar{\boldsymbol{\alpha}}) &\geq f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}) - \langle \Phi_n^*\boldsymbol{\alpha}, \bar{\boldsymbol{w}} - \boldsymbol{w}\rangle \\ &= f(\bar{\boldsymbol{w}}) - f(\boldsymbol{w}) - \langle \nabla f(\boldsymbol{w}), \bar{\boldsymbol{w}} - \boldsymbol{w}\rangle, \\ &= \frac{1}{p}\|\bar{\boldsymbol{w}}\|_p^p - \frac{1}{p}\|\boldsymbol{w}\|_p^p - \langle J_p(\boldsymbol{w}), \bar{\boldsymbol{w}} - \boldsymbol{w}\rangle.\end{aligned}$$

Now, since $1 < p < 2$, it follows from [2, Corollary 2.6.1] that

$$\frac{1}{p}\|\bar{\boldsymbol{w}}\|_p^p - \frac{1}{p}\|\boldsymbol{w}\|_p^p - \langle J_p(\boldsymbol{w}), \bar{\boldsymbol{w}} - \boldsymbol{w}\rangle \geq \frac{C_p}{\big(\|\bar{\boldsymbol{w}}\|_p + \|\boldsymbol{w}\|_p\big)^{2-p}}\|\bar{\boldsymbol{w}} - \boldsymbol{w}\|_p^2,$$

for some constant $C_p > 0$ that depends only on p. Therefore, by the definition of the duality map,

$$\|\boldsymbol{w}\|_p = \|J_q(\Phi_n^*\boldsymbol{\alpha})\|_p = \big(\|\Phi_n^*\boldsymbol{\alpha}\|_q^q\big)^{1/p} \leq q^{1/p}\big(\Lambda(\boldsymbol{\alpha}) + \gamma\|\boldsymbol{\xi}\|_1\big)^{1/p},$$

where $\xi_i = \inf L^*(y_i,\cdot)$; and similarly for $\|\bar{\boldsymbol{w}}\|_p$. Then the statement follows. □

Proof of Theorem 2.3. Since for the least squares loss we have $\xi_i = -(1/2)y_i^2$, it follows from Proposition 3.2 that for every $m \in \mathbb{N}$,

$$\|\boldsymbol{w}_m - \bar{\boldsymbol{w}}\|_p^2 \leq \frac{\big[(2^p q)\big(\Lambda(\boldsymbol{\alpha}_m) + (\gamma/2)\|\boldsymbol{y}\|_2^2\big)\big]^{(2-p)/p}}{C_p} \cdot \big(\Lambda(\boldsymbol{\alpha}_m) - \min\Lambda\big).$$

Now it remains to prove that, $\inf_m \lambda_m > 0$ and that

$$(\forall\, m \in \mathbb{N}) \quad \Lambda(\boldsymbol{\alpha}_{m+1}) - \min\Lambda \leq \big(1 - (2/\gamma)\lambda_m(1-\delta)\big)\big(\Lambda(\boldsymbol{\alpha}_m) - \min\Lambda\big). \tag{33}$$

First of all, since $q > 2$, the gradient of Λ is Lipschitz continuous on bounded sets. Therefore, Proposition 3.15 in [13] yields that $\inf_m \lambda_m > 0$. Now, because of the linesearch rule we have that

$$\Lambda(\boldsymbol{\alpha}_{m+1}) \leq \Lambda(\boldsymbol{\alpha}_m) - \lambda_m(1-\delta)\|\nabla\Lambda(\boldsymbol{\alpha}_m)\|_2^2$$

and, since Λ is strongly convex with modulus $1/\gamma$, we have

$$\Lambda(\boldsymbol{\alpha}_m) - \Lambda(\bar{\boldsymbol{\alpha}}) \leq \frac{\gamma}{2}\|\nabla\Lambda(\boldsymbol{\alpha}_m)\|_2^2.$$

All together the two inequalities above gives

$$\Lambda(\boldsymbol{\alpha}_{m+1}) \leq \Lambda(\boldsymbol{\alpha}_m) - (2/\gamma)\lambda_m(1-\delta)\big(\Lambda(\boldsymbol{\alpha}_m) - \Lambda(\bar{\boldsymbol{\alpha}})\big).$$

Adding $\Lambda(\bar{\boldsymbol{\alpha}})$ to both sides, (33) follows and hence the statement. □

B.2 The function Banach space associated to a tensor kernel

In this section we make explicit the space associated to tensor kernels. We assume that $\operatorname{span}(\Phi(\mathbb{R}^d))$ is dense in $\ell^q(\mathbb{K})$ – which is equivalent to requiring that the functions $(\phi_k)_{k\in\mathbb{K}}$ are ℓ^p point-wise independent. Then, we can associate to the feature map Φ the Banach function space [20]

$$\mathcal{B} = \big\{\langle \boldsymbol{w}, \Phi(\cdot)\rangle \,\big|\, \boldsymbol{w} \in \ell^p(\mathbb{K})\big\},\ \ \|\langle \boldsymbol{w}, \Phi(\cdot)\rangle\|_{\mathcal{B}} = \|\boldsymbol{w}\|_p. \tag{34}$$

Note that if $\boldsymbol{\alpha} \in \mathbb{R}^n$, $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$, and we set $\boldsymbol{w} = J_q\big(\sum_{i=1}^n \alpha_i \Phi(\boldsymbol{x}_i)\big)$, then, as in (14), we have

$$\langle \boldsymbol{w}, \Phi(\cdot)\rangle = \sum_{i_1,\ldots,i_{q-1}=1}^{n} K(\boldsymbol{x}_{i_1}, \cdots, \boldsymbol{x}_{i_{q-1}}, \cdot)\alpha_{i_1}\cdots\alpha_{i_{q-1}}, \tag{35}$$

and

$$\|\langle \boldsymbol{w}, \Phi(\cdot)\rangle\|_{\mathcal{B}} = \bigg(\sum_{i_1,\ldots,i_q=1}^{n} K(\boldsymbol{x}_1, \ldots, \boldsymbol{x}_q)\alpha_{i_1}\cdots\alpha_{i_q}\bigg)^{1/p}, \tag{36}$$

and the functions (35) are dense in $\mathcal{B}$. Moreover, setting $\Phi^* = J_q \circ \Phi\colon \mathbb{R}^d \to \ell^p(\mathbb{K})$, if $\operatorname{span}(\Phi^*(\mathbb{R}^d))$ is also dense in $\ell^p(\mathbb{K})$, then its associated function Banach space $\mathcal{B}^*$ (defined similarly to (34)) is the topological dual of $\mathcal{B}$ and the following reproducing property holds

$$K_{\boldsymbol{x}}\colon \boldsymbol{x}' \to K(\boldsymbol{x}', \ldots, \boldsymbol{x}', \boldsymbol{x}) \in \mathcal{B}^* \text{ and } \langle f, K_{\boldsymbol{x}}\rangle = f(\boldsymbol{x}).$$

For the case of infinite dimensional power series tensor kernels, which includes the exponential tensor kernels considered here, the density assumptions on $\operatorname{span}(\Phi(\mathbb{R}^d))$ and $\operatorname{span}(\Phi^*(\mathbb{R}^d))$ holds, hence the corresponding Banach space can be described through the equations (35) and (37).

B.3 The dual algorithm for general loss function and any $p \in]1, 2[$

Proof of Theorem 3.6. Since φ_1 is smooth with a locally Lipschitz continuous gradient we can apply Theorem 3.2 and Proposition 3.5 in [13] and get $\inf_m \lambda_m > 0$, $\boldsymbol{\alpha}_m \to \bar{\boldsymbol{\alpha}}$ and $\Lambda(\boldsymbol{\alpha}_m) - \Lambda(\bar{\boldsymbol{\alpha}}) = o(1/m)$. Then, by Proposition 3.2, we have $\|\boldsymbol{w}_m - \bar{\boldsymbol{w}}\|_p \leq o(1/\sqrt{m})$. Now suppose that Λ is μ-strongly convex. We will rely on Proposition 2 in [5]. Then, strong convexity of Λ yields

$$\frac{\mu}{2}\|\boldsymbol{\alpha}_m - \bar{\boldsymbol{\alpha}}\|^2 \leq \Lambda(\boldsymbol{\alpha}_m) - \Lambda(\bar{\boldsymbol{\alpha}})$$

for some constant $\mu > 0$. So equation (3.8) in Proposition 2 in [5] holds. Moreover, defining

$$-D_{\lambda_m}(\boldsymbol{\alpha}_m) := \varphi_2(\boldsymbol{\alpha}_{m+1})\big) - \varphi_2(\boldsymbol{\alpha}_m)\big) + \langle \boldsymbol{\alpha}_{m+1} - \boldsymbol{\alpha}_m, \nabla\varphi_1(\boldsymbol{\alpha}_m)\rangle,$$

by the definition of λ_m, and Proposition 3.8 and Proposition 3.9 in [13], we have

$$\frac{\|\boldsymbol{\alpha}_{m+1} - \boldsymbol{\alpha}_m\|^2}{\lambda_m} \leq D_{\lambda_m}(\boldsymbol{\alpha}_m) \quad \text{and} \quad \Lambda(\boldsymbol{\alpha}_{m+1}) - \Lambda(\boldsymbol{\alpha}_m) \leq -(1-\delta)D_{\lambda_m}(\boldsymbol{\alpha}_m). \tag{37}$$

Then, since $\inf_m \lambda_m > 0$ we can proceed as in the proof of Proposition 2 in [5] and prove that $\Lambda(\boldsymbol{\alpha}_m)$ converge linearly to $\Lambda(\bar{\boldsymbol{\alpha}})$. Finally, using Proposition 3.2 the linear convergence of $\boldsymbol{w}_m$ follows. Note that Example B.2 shows that if L is the least square loss or the logistic loss, then Λ is strongly convex. □