6. Supplement

6.1. Subspace-Valued Maps

Proposition 6.1. *Let S denote a quasilinear subspace-valued map. Then*

$$S(\alpha x) = \alpha S(x)$$

and

$$\alpha S(x) \subseteq S(\alpha x + \beta y) + \beta S(y)$$

for every $x, y \in \mathcal{H}$, $\alpha, \beta \in \mathbb{R}$.

Proof. The case $\alpha = 0$ follows directly from the definition. If $\alpha \neq 0$, applying quasilinearity with $\beta \leftarrow 0$ we obtain that $S(\alpha x) \subseteq \alpha S(x)$ and $S(x) \subseteq \frac{1}{\alpha} S(\alpha x)$. From these (2) follows. (3) follows from (2) and the definition applied to the difference of $\alpha x + \beta y$ and βy. □

Lemma 6.1. *Let S be a quasilinear and idempotent subspace-valued map. Then, for every $m \in \mathbb{N}$ and every set $\{x_i : i \in \mathbb{N}_m\} \subseteq \mathcal{H}$, it holds that*

$$S\left(\sum_{i=1}^{m} S(x_i)\right) \subseteq \sum_{i=1}^{m} S(x_i).$$

Proof. From quasilinearity and idempotence we obtain that $S(S(x) + y) \subseteq S(x) + S(y)$ for every $x, y \in \mathcal{H}$. The assertion then follows by induction. □

6.2. Loss Functions Which Lead to Orthomonotonicity

Lemma 6.2. *Assume that V is a regression loss function. Then, for every $p \in \mathbb{N}$, there exist output data $\{y_i : i \in \mathbb{N}_{2p}\} \subset \mathbb{R}$ and a function $f_u : \mathbb{R}^p \to \mathbb{R} \cup \{+\infty\}$ satisfying the hypothesis of Theorem 3.1, Part 2, such that the error functional*

$$w \mapsto f(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle, \langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) \ ,$$

with f defined by (14) for $m = 2p$, equals the error functional

$$w \mapsto f_u(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) \ .$$

Proof. By the hypothesis on ϕ, it follows that the set of minimizers $\mathcal{M} = \{t \in \mathbb{R} : \phi(t) = \phi(0)\}$ is closed and bounded. If ϕ is uniquely minimized at zero, that is $\mathcal{M} = \{0\}$, we have immediately that f satisfies the hypothesis of part 2 of Theorem 3.1 if we select the output data vector to be nonzero. Otherwise, let $\alpha = \min \mathcal{M}$ and $\beta = \max \mathcal{M}$, and consider the error function

$$f_u(z) = \sum_{i=1}^{p} V_u(z_i, y_i) \ ,$$

$$V_u(z, y) = \phi(z - y) + \phi(z - y - \beta + \alpha).$$

Observe that the function $V_u(z, y)$ is lower semicontinuous and has bounded sublevel sets. Moreover, it is uniquely minimized for $z = y + \beta$, since

- if $z = y + \beta$, then $V_u(z, y) = \phi(\beta) + \phi(\alpha) = 2\phi(0)$,
- if $z > y + \beta$, then $V_u(z, y) \geq \phi(z - y) + \phi(0) > 2\phi(0)$,
- if $z < y + \beta$, then $V_u(z, y) \geq \phi(0) + \phi(z - y - \beta + \alpha) > 2\phi(0)$.

The error function f_u is lower semicontinuous with bounded sublevel sets and uniquely minimized for $z_i = y_i + \beta$, thus satisfying the hypothesis of part 2 of Theorem 3.1, provided that $y_i \neq -\beta$ for some $i \in \mathbb{N}_p$. Finally, observe that

$$\begin{aligned} &f(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle, \langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) \\ &\quad = f_u(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) \qquad \forall w \in \mathcal{H} \end{aligned}$$

if we choose $y_{p+i} = y_i + (\beta - \alpha)$ for all $i \in \mathbb{N}_p$. □

Lemma 6.3. *Assume that V is a regular binary classification loss function. Then, for every $p \in \mathbb{N}$, there exist output data $\{y_i : i \in \mathbb{N}_{2p}\} \subseteq \{-1, +1\}$ and a function $f_u : \mathbb{R}^p \to \mathbb{R}$ satisfying the hypothesis of Theorem 3.1, Part 2, such that the error functional*

$$w \mapsto f(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle, \langle w, \alpha w_1\rangle, \ldots, \langle w, \alpha w_p\rangle)) \ ,$$

with f defined by (14) for $m = 2p$, equals the error functional

$$w \mapsto f_u(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) \ .$$

Proof. For any $p \in \mathbb{N}$ and $y \in \{-1, +1\}^p$, consider the error function

$$f_u(z) = \sum_{i=1}^{p} \psi_\alpha(z_i y_i) \ .$$

In view of the hypothesis on ϕ, the function f_u satisfies the hypothesis of part 2 of Theorem 3.1. Moreover, observe that

$$\begin{aligned} &f(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle, \langle w, \alpha w_1\rangle, \ldots, \langle w, \alpha w_p\rangle) = \\ &\quad = f_u(\langle w, w_1\rangle, \ldots, \langle w, w_p\rangle) \quad \forall w \in \mathcal{H} \end{aligned}$$

if we choose $y_{p+i} = -y_i$ for all $i \in \mathbb{N}_p$. □

6.3. Properties of Orthomonotone Functions

Proposition 6.2. *Let $a \in \mathcal{H}$ and $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ orthomonotone with respect to the map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. If S is quasilinear then the function $x \mapsto \Omega(x + a)$ is orthomonotone with respect to the map $x \mapsto S(x) + S(a)$.*

Proof. Quasilinearity implies that $S(x+a) \subseteq S(x)+S(a)$, for every $x \in \mathcal{H}$, and the assertion follows from the definition of orthomonotonicity. □

Proposition 6.3. *Let $\Omega_1 : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be orthomonotone with respect to a map $S_1 : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ and $\Omega_2 : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be orthomonotone with respect to a map $S_2 : \mathcal{H} \to \mathcal{V}(\mathcal{H})$. Also let $h : (\mathbb{R} \cup \{+\infty\})^2 \to \mathbb{R} \cup \{+\infty\}$ be elementwise nondecreasing, that is, $h(a', b') \geq h(a, b)$ whenever $a' \geq a$ and $b' \geq b$. Then the function $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$,*

$$\Omega(w) = h\left(\Omega_1(w), \Omega_2(w)\right) \qquad \forall w \in \mathcal{H},$$

is orthomonotone with respect to the map $S_1 + S_2$.

Proof. The assertion follows by combining the orthomonotonicities of Ω_1, Ω_2 with the fact that if $w \in \mathcal{H}$, $p \in (S_1 + S_2)(w)^\perp$ then $p \in S_1(w)^\perp \cap S_2(w)^\perp$. □

Note that such operations enlarge, in general, the class of orthomonotone functions, since the image of $S_1 + S_2$ at any fixed point contains those of S_1, S_2. Thus, Proposition 3.3 can be used to obtain representer theorems for new penalties, based on known representer theorems. However, this technique may not provide the complete class of orthomonotone penalties for the sum of the maps. An example of this is the multitask representer theorem (see Example 4.2) which yields a class of penalties of the form $h(W^\top W)$. Applying the "classical" representer theorem on the space of matrices $\mathbf{M}_{d,n}$ yields the subclass of the form $h(\|W\|^2_{Frob})$. Considering the maps $S_{ij}(X) = \operatorname{span}\{XE_{ij}\}$, $i \in \mathbb{N}_d, j \in \mathbb{N}_n$, each of which corresponds to the class of monotone functions of $\|w_i\|$, we could apply Proposition 3.3 and obtain the class of penalties $h(\|w_1\|^2, \ldots, \|w_n\|^2)$, which is strictly nested between the previous two classes.

Proposition 6.4. *Let $\Omega : \mathcal{H} \to \mathbb{R} \cup \{+\infty\}$ be orthomonotone with respect to a map $S : \mathcal{H} \to \mathcal{V}(\mathcal{H})$ and let $T \in \mathscr{L}(\mathcal{H})$ be a continuous operator. Then the function $\Omega \circ T$ is orthomonotone with respect to $T^* \circ S \circ T$.*

Proof. Let $x \in \mathcal{H}$, $y \in (T^* \circ S \circ T)(x)^\perp$. Then $Ty \in S(Tx)^\perp$ and, by orthomonotonicity of Ω, we obtain that $\Omega(Tx + Ty) \geq \Omega(Tx)$. □

6.4. Geometric Interpretation of Orthomonotonicity

In the case of convex regularizers Ω, orthomonotonicity can be rephrased as the property that the affine subspace $x + S(x)^\perp$ is tangent to the contour passing through x. Figure 1 illustrates this for both a nonsmooth Ω (top) and a smooth Ω (bottom). Let Λ be the convex cone tangent to the contour at x. In the top plot, Λ is delimited by the dashed lines. In the smooth case, Λ is a halfspace. In both cases, orthomonotonicity means that $x + S(x)^\perp$ is contained in $\mathcal{H} \setminus \Lambda$.

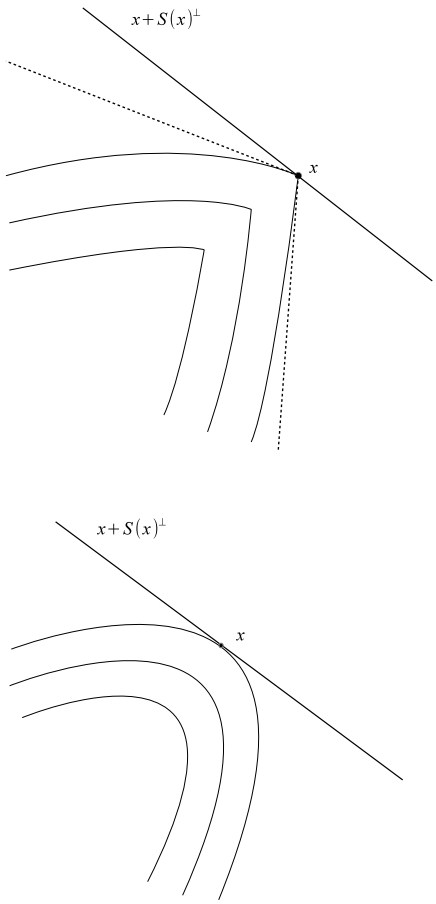

Figure 1. Interpretation of orthomonotonicity for convex functions Ω. The curves depict contours of Ω and the dashed lines the boundary of the tangent cone.

6.5. Alternative Regularization Techniques

Example 4.1 refers to the representer theorem we call "classical", that is, to the case of regularizers which are nondecreasing functions of the Hilbertian norm. In particular, the classical theorem applies to the widely used method of *Tikhonov regularization*. However alternative ways to formulate the optimization problem, which are essentially equivalent to Tikhonov regularization, have also been used. Let us consider the following optimization problems,

$$\min\{\mathcal{E}_y(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) + \gamma \|w\|^2 : w \in \mathcal{H}\} \qquad (Tikhonov)$$

$$\min\{\mathcal{E}_y(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) : \|w\| \leq r, w \in \mathcal{H}\} \quad (Ivanov)$$

$$\min\{\|w\|^2 : \mathcal{E}_y(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) \leq \eta^2, w \in \mathcal{H}\} \quad (Phillips)$$

$$\min\{\mathcal{E}_y(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) + \left(\tfrac{\eta}{r}\right)^2 \|w\|^2 : \mathcal{E}_y(\langle w, w_1\rangle, \ldots, \langle w, w_m\rangle) \leq \eta^2, \|w\| \leq r, w \in \mathcal{H}\} \quad (Miller)$$

where γ, r, η are positive regularization parameters and $\mathcal{E}_y : z \mapsto \frac{1}{m}\|z - y\|^2$ with $y \in \mathbb{R}^m$ a fixed output vector.

It is easy to see that each of the above optimization problems can be rephrased as the minimization of a functional J of the form (4) with appropriate choices of f and Ω. For example, Ivanov regularization is obtained with the choice $f = \mathcal{E}_y$, $\Omega : w \mapsto \begin{cases} 0 & \text{if } \|w\| \leq r \\ +\infty & \text{otherwise} \end{cases}$. Applying the first part of Theorem 3.1 yields a representer theorem for all of the above regularization problems, which is a known fact (Gnecco & Sanguineti, 2010; Schölkopf & Smola, 2002).

But in addition, part 2 of Theorem 3.1 and Lemma 3.1 imply the *necessity* of radial regularizers in the Tikhonov and Ivanov formulations. That is, if we replaced $\|\cdot\|$ in the Tikhonov or Ivanov problems with a penalty Ω satisfying the assumptions of Theorem 3.1, Part 2, and assumed that the classical representer theorem holds for any choice of data $w_1, \ldots, w_m \in \mathcal{H}$ and any $\gamma > 0$, then Ω would have to be radial if $\dim \mathcal{H} \geq 2$ (see Example 4.1). In fact, for the Ivanov formulation it suffices to assume that the representer theorem holds for a *single* value of $r > 0$ (since r is a parameter of the regularizer Ω). In contrast, for the Phillips formulation it remains an open question whether radial functions are the only regularizers yielding the classical representer theorem. In this case, the error term f appears in the constraint and hence does not admit a unique minimizer as required in part 2 of Theorem 3.1. Regarding the Miller formulation, part 2 of Theorem 3.1 does not apply directly since η and r are parameters of f and Ω, respectively, but would apply with the inclusion of a free parameter $\gamma > 0$ multiplying Ω.

Finally, let us remark that the above ideas can be extended in a straightforward way to generalized representer theorems. In other words, representer theorems can be obtained for regularization problems of Tikhonov, Ivanov, Phillips or Miller type, in which the regularizer Ω is orthomonotone with respect to an arbitrary regular quasilinear map S (such as the examples of Section 4).

6.6. Tensor Learning

A representer theorem can also be derived for tensor learning problems. Consider a regularization problem for learning a 3-way tensor

$$\min\{f(\langle W, W_1\rangle, \ldots, \langle W, W_m\rangle) + \gamma_1 \Omega_1(\mathrm{Mat}_1(W)) + \gamma_2 \Omega_2(\mathrm{Mat}_2(W)) + \gamma_3 \Omega_3(\mathrm{Mat}_3(W)) : W \in \mathbb{R}^{d_1 \times d_2 \times d_3}\}. \quad (16)$$

Here Mat_i is the operator that maps a tensor to its i-th matrix unfolding and $\Omega_1 : \mathbf{M}_{d_1, d_2 d_3} \to \mathbb{R} \cup \{+\infty\}$, $\Omega_2 : \mathbf{M}_{d_2, d_1 d_3} \to \mathbb{R} \cup \{+\infty\}$, $\Omega_3 : \mathbf{M}_{d_3, d_1 d_2} \to \mathbb{R} \cup \{+\infty\}$ are functions of the form

$$\Omega_i(X) = h_i(X^\top X)$$

with h_i a matrix nondecreasing function. Examples of such penalties Ω_i are *spectral functions* of the matricizations, *weighted spectral functions* of the matricizations, group Lasso type *mixed* $(2, p)$ *norms* of the matricizations etc. Let $T_i = \mathrm{Mat}_i$ and S_i similar to Example 2.3, for $i = 1, 2, 3$. The case of spectral penalties on matricizations has been proposed and studied recently – see (Signoretto et al., 2013) and references therein.

Applying Propositions 3.3 and 3.4, we obtain that the penalty in (16) is orthomonotone with respect to the map $S' = T_1^* \circ S_1 \circ T_1 + T_2^* \circ S_2 \circ T_2 + T_3^* \circ S_3 \circ T_3$. Since $T_i^* = \mathrm{Mat}_i^{-1}$, the map S' is idempotent and hence regular quasilinear. Thus we obtain the following representer theorem.

Corollary 6.1. *If problem* (16) *admits a minimizer then there exists a minimizer* $\hat{W}$ *of the form*

$$\hat{W} = \sum_{i=1}^m \mathrm{Mat}_1^{-1}\left(\mathrm{Mat}_1(W_i) C_i^{(1)}\right) + \sum_{i=1}^m \mathrm{Mat}_2^{-1}\left(\mathrm{Mat}_2(W_i) C_2^{(i)}\right) + \sum_{i=1}^m \mathrm{Mat}_3^{-1}\left(\mathrm{Mat}_3(W_i) C_3^{(i)}\right)$$

for some $C_i^{(1)} \in \mathbf{M}_{d_2 d_3}, C_i^{(2)} \in \mathbf{M}_{d_1 d_3}, C_i^{(3)} \in \mathbf{M}_{d_1 d_2}$, $\forall i \in \mathbb{N}_m$.

Clearly the result generalizes to tensors of any order. A related representer theorem for the special case of spectral penalties has recently appeared in (Signoretto et al., 2013).