Appendix for Mixed-norm Regularization for Brain Decoding

Rémi Flamary, Nisrine Jrad, Ronald Phlypo, Marco Congedo, Alain Rakotomamonjy
LITIS, EA 4108 - INSA / Université de Rouen
Avenue de l'Université - 76801 Saint-Etienne du Rouvray Cedex
firstname.lastname@insa-rouen.fr

APPENDIX

A. Proof of Lipschitz gradient of the squared Hinge loss

Given the training examples $\{\mathbf{x}_i, y_i\}$, the squared Hinge loss is written as :

$$J = \sum_{i=1}^{n} \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})^2$$

and its gradient is :

$$\nabla_{\mathbf{w}} J = -2 \sum_i \mathbf{x}_i y_i \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$$

The squared Hinge loss is gradient Lipschitz if there exists a constant L such that:

$$\|\nabla J(\mathbf{w}_1) - \nabla J(\mathbf{w}_2)\|_2 \leq L \|\mathbf{w}_1 - \mathbf{w}_2\|_2 \quad \forall \mathbf{w}_1, \mathbf{w}_2 \in \mathbb{R}^d.$$

The proof essentially relies on showing that $\mathbf{x}_i y_i \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$ is Lipschitz itself *i.e* there exists $L' \in \mathbb{R}$ such that

$$\|\mathbf{x}_i y_i \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w}_1) - \mathbf{x}_i y_i \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w}_2)\| \leq L' \|\mathbf{w}_1 - \mathbf{w}_2\|$$

Now let us consider different situations. For a given $\mathbf{w}_1$ and $\mathbf{w}_2$, if $1 - \mathbf{x}_i^T \mathbf{w}_1 \leq 0$ and $1 - \mathbf{x}_i^T \mathbf{w}_2 \leq 0$, then the left hand side is equal to 0 and any L' would satisfy the inequality. If $1 - \mathbf{x}_i^T \mathbf{w}_1 \leq 0$ and $1 - \mathbf{x}_i^T \mathbf{w}_2 \geq 0$, then the left hand side (lhs) is

$$\begin{aligned} lhs &= \|\mathbf{x}_i\|_2 (1 - \mathbf{x}_i^\top \mathbf{w}_2) \\ &\leq \|\mathbf{x}_i\|_2 (\mathbf{x}_i^\top \mathbf{w}_1 - \mathbf{x}_i^\top \mathbf{w}_2) \\ &\leq \|\mathbf{x}_i\|_2^2 \|\mathbf{w}_1 - \mathbf{w}_2\|_2 \end{aligned} \quad (1)$$

A similar reasoning yields to the same bound when $1 - \mathbf{x}_i^T \mathbf{w}_1 \geq 0$ $1 - \mathbf{x}_i^T \mathbf{w}_1 \leq 0$ and $1 - \mathbf{x}_i^T \mathbf{w}_2 \geq 0$ and $1 - \mathbf{x}_i^T \mathbf{w}_2 \geq 0$. Thus, $\mathbf{x}_i y_i \max(0, 1 - y_i \mathbf{x}_i^\top \mathbf{w})$ is Lipschitz with a constant $\|\mathbf{x}_i\|^2$. Now, we can conclude the proof by stating that $\nabla_{\mathbf{w}} J$ is Lipschitz as it is a sum of Lipschitz function and the related constant is $\sum_{i=1}^n \|\mathbf{x}_i\|_2^2$.

This work was partly supported by the FP7-ICT Programme of the European Community, under the PASCAL2 Network of Excellence, ICT- 216886, by the French ANR Project ASAP ANR-09-EMER-001, OpenVibe2, GazeEEg, and the INRIA ARC MABI.

B. Lipschitz gradient for the multi-task learning problem

For the multi-task learning problem, we want to prove that the function

$$\sum_{t=1}^{m} \sum_{i=1}^{n} L(y_{i,t}, \mathbf{x}_{i,t}^\top \mathbf{w}_t + \mathbf{b}_t) + \lambda_s \sum_{t=1}^{m} \|\mathbf{w}_t - \frac{1}{m} \sum_{j=1}^{m} \mathbf{w}_j\|_2^2$$

is gradient Lipschitz, $L(\cdot, \cdot)$ being the square Hinge loss. From the above results, it is easy to show that the first term is gradient Lipschitz as the sum of gradient Lipschitz functions.

Now, we also show that the similarity term

$$\sum_t \|\mathbf{w}_t - \frac{1}{m} \sum_{j=1}^{m} \mathbf{w}_j\|_2^2$$

is also gradient Lipschitz.

This term can be expressed as

$$\begin{aligned} \|\mathbf{w}_t - \frac{1}{m} \sum_{j=1}^{m} \mathbf{w}_j\|_2^2 &= \sum_t \langle \mathbf{w}_t, \mathbf{w}_t \rangle - \frac{1}{m} \sum_{i,j=1}^{m} \langle \mathbf{w}_i, \mathbf{w}_j \rangle \\ &= \mathbf{w}^\top \mathbf{M} \mathbf{w} \end{aligned}$$

where $\mathbf{w}^\top = [\mathbf{w}_1^\top, \ldots, \mathbf{w}_m^\top]$ is the vector of all classifier parameters and $\mathbf{M} \in \mathbb{R}^{md \times md}$ is the Hessian matrix of the similarity regularizer of the form

$$\mathbf{M} = \mathbf{I} - \frac{1}{m} \sum_{t=1}^{m} \mathbf{D}_t$$

with $\mathbf{I}$ the identity matrix and $\mathbf{D}_t$ a block matrix with $\mathbf{D}_t$ a $(t-1)$-diagonal matrix where each block is an

identity matrix $\mathbf{I}$ with appropriate circular shift. $\mathbf{D}_t$ is thus a $(t-1)$ row-shifted version of $\mathbf{I}$.

Once we have this formulation, we can use the fact that a function f is gradient Lipschitz of constant L if the largest eigenvalue of its Hessian is bounded by L on its domain [1]. Hence, since we have

$$\|\mathbf{M}\|_2 \leq \|\mathbf{I}\|_2 + \frac{1}{m}\sum_{t=1}^{m}\|\mathbf{D}_t\|_2 = 2$$

the Hessian matrix of the similarity term $2 \cdot \mathbf{M}$ has consequently bounded eigenvalues. This concludes the proof that the function $\mathbf{w}^\top \mathbf{M}\mathbf{w}$ is gradient Lipschitz continuous.

C. Proximal operators

1) ℓ_1 norm: the proximal operator of the ℓ_1 norm is defined as :

$$\text{prox}_{\lambda\|\mathbf{x}\|_1}(\mathbf{u}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 + \lambda\|\mathbf{x}\|_1$$

and has the following closed-form solution for which each component is

$$[\text{prox}_{\lambda\|\mathbf{x}\|_1}(\mathbf{u})]_i = \text{sign}(u_i)(|u_i| - \lambda)_+$$

2) $\ell_1 - \ell_2$ norm: the proximal operator of the $\ell_1 - \ell_2$ norm is defined as :

$$\text{prox}_{\lambda\sum_{g\in\mathcal{G}}\|\mathbf{x}_g\|_2}(\mathbf{u}) = \arg\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 + \lambda\sum_{g\in\mathcal{G}}\|\mathbf{x}_g\|_2$$

the minimization problem can be decomposed into several ones since the indices g are separable. Hence, we can just focus on the problem

$$\min_{\mathbf{x}} \frac{1}{2}\|\mathbf{x}-\mathbf{u}\|_2^2 + \lambda\|\mathbf{x}\|_2$$

which minimizer is

$$\begin{cases} 0 & \text{if } \|\mathbf{u}\|_2 \leq \lambda \\ (1 - \frac{\lambda}{\|\mathbf{u}\|_2})\mathbf{u} & \text{otherwise} \end{cases}$$

REFERENCES

[1] D. Bertsekas, *Nonlinear programming*. Athena scientific, 1999.