Supplementary Materials
Localized Lasso for High-Dimensional Regression

Makoto Yamada[1,2], Koh Takeuchi[3], Tomoharu Iwata[3], John Shawe-Taylor[4], Samuel Kaski[5]
[1]RIKEN AIP, Japan
[2]JST, PRESTO
[3]NTT Communication Science Laboratories, Japan
[4]University College London, UK
[5]Aalto University, Finland
makoto.yamada@riken.jp, {takeuchi.koh,iwata.tomoharu}@lab.ntt.co.jp
j.shawe-taylor@ucl.ac.uk, samuel.kaski@aalto.fi

Propositions used for deriving Eq. (4) in main paper

Proposition 1 *Under $r_{ij} \geq 0$, $r_{ij} = r_{ji}$, $r_{ii} = 0$, we have*

$$\frac{\partial}{\partial \text{vec}(\boldsymbol{W})} \sum_{i,j=1}^{n} r_{ij} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 = 2\boldsymbol{F}_g \text{vec}(\boldsymbol{W}),$$

where

$$\boldsymbol{F}_g = \boldsymbol{I}_d \otimes \boldsymbol{C},$$
$$[\boldsymbol{C}]_{i,j} = \begin{cases} \sum_{j'=1}^{n} \frac{r_{ij'}}{\|\boldsymbol{w}_i - \boldsymbol{w}'_j\|_2} - \frac{r_{ij}}{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2} & (i = j) \\ \frac{-r_{ij}}{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2} & (i \neq j) \end{cases}.$$

Proof: Under $r_{ij} \geq 0$, $r_{ij} = r_{ji}$, $r_{ii} = 0$, the derivative of the network regularization term with respect to $\boldsymbol{w}_k$ is given as

$$\begin{aligned}
\frac{\partial}{\partial \boldsymbol{w}_k} \sum_{i,j=1}^{n} r_{ij} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 &= \sum_{i=1}^{n} r_{ik} \frac{\boldsymbol{w}_k - \boldsymbol{w}_i}{\|\boldsymbol{w}_k - \boldsymbol{w}_i\|_2} + \sum_{j=1}^{n} r_{kj} \frac{\boldsymbol{w}_k - \boldsymbol{w}_j}{\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2} \\
&= \boldsymbol{w}_k \left(\sum_{i=1}^{n} \frac{r_{ik}}{\|\boldsymbol{w}_k - \boldsymbol{w}_i\|_2} + \sum_{j=1}^{n} \frac{r_{kj}}{\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2} \right) \\
&\quad - \sum_{i=1}^{n} \frac{r_{ik}}{\|\boldsymbol{w}_k - \boldsymbol{w}_i\|_2} \boldsymbol{w}_i - \sum_{j=1}^{n} \frac{r_{kj}}{\|\boldsymbol{w}_j - \boldsymbol{w}_k\|_2} \boldsymbol{w}_j \\
&= 2 \left(\boldsymbol{w}_k \sum_{i=1}^{n} \frac{r_{ik}}{\|\boldsymbol{w}_k - \boldsymbol{w}_i\|_2} - \sum_{i=1}^{n} \frac{r_{ik}}{\|\boldsymbol{w}_k - \boldsymbol{w}_i\|_2} \boldsymbol{w}_i \right).
\end{aligned}$$

Thus,

$$\frac{\partial}{\partial \boldsymbol{W}} \sum_{i,j=1}^{n} r_{ij} \|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 = 2\boldsymbol{C}\boldsymbol{W},$$

where $\boldsymbol{W} = [\boldsymbol{w}_1, \ldots, \boldsymbol{w}_n]^\top \in \mathbb{R}^{n\times d}$*. Since* $\mathrm{vec}(\boldsymbol{C}\boldsymbol{W}\boldsymbol{I}_d) = (\boldsymbol{I}_d \otimes \boldsymbol{C})\mathrm{vec}(\boldsymbol{W})$*, we have*

$$\frac{\partial}{\partial \mathrm{vec}(\boldsymbol{W})} \sum_{i,j=1}^{n} r_{ij}\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2 = 2(\boldsymbol{I}_d \otimes \boldsymbol{C})\mathrm{vec}(\boldsymbol{W}),$$

where $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ *is the identity matrix and* $\mathrm{vec}(\cdot)$ *is the vectorization operator.* □

Proposition 2

$$\frac{\partial}{\partial \mathrm{vec}(\boldsymbol{W})} \sum_{i=1}^{n} \|\boldsymbol{w}_i\|_1^2 = 2\boldsymbol{F}_e \mathrm{vec}(\boldsymbol{W}),$$

where

$$[\boldsymbol{F}_e]_{\ell,\ell} = \sum_{i=1}^{n} \frac{I_{i,\ell}\|\boldsymbol{w}_i\|_1}{[\mathrm{vec}(|\boldsymbol{W}|)]_\ell}.$$

Hence, $I_{i,\ell} \in \{0,\ 1\}$ *are the group index indicators:* $I_{i,\ell} = 1$ *if the* ℓ*-th element* $[\mathrm{vec}(\boldsymbol{W})]_\ell$ *belongs to group* i *(i.e.,* $[\mathrm{vec}(\boldsymbol{W})]_\ell$ *is the element of* $\boldsymbol{w}_i$*), otherwise* $I_{i,\ell} = 0$*.*

Propositions and lemmas used for deriving Theorem 1 in main paper

Proposition 3 *Under* $r_{ij} \geq 0$*,* $r_{ij} = r_{ji}$*,* $r_{ii} = 0$*, we have*

$$\mathrm{vec}(\boldsymbol{W})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W}) = \sum_{i,j=1}^{n} r_{ij} \frac{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2^2}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2},$$

where

$$\boldsymbol{F}_g^{(t)} = \boldsymbol{I}_d \otimes \boldsymbol{C}^{(t)},$$
$$[\boldsymbol{C}^{(t)}]_{i,j} = \begin{cases} \sum_{j'=1}^{n} \frac{r_{ij'}}{\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_{j'}^{(t)}\|_2} - \frac{r_{ij}}{\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} & (i = j) \\ \frac{-r_{ij}}{\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} & (i \neq j) \end{cases}.$$

Proof:

$$\begin{aligned} &\sum_{i,j=1}^{n} r_{ij} \frac{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2^2}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} \\ &= \sum_{i=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{w}_i \sum_{j=1}^{n} \frac{r_{ij}}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} + \sum_{j=1}^{n} \boldsymbol{w}_j^\top \boldsymbol{w}_j \sum_{i=1}^{n} \frac{r_{ij}}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} - 2\sum_{i=1}^{n}\sum_{j=1}^{n} \boldsymbol{w}_i^\top \boldsymbol{w}_j \frac{r_{ij}}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} \\ &= \mathrm{tr}(\boldsymbol{W}^\top \boldsymbol{C}^{(t)} \boldsymbol{W}) \\ &= \mathrm{vec}(\boldsymbol{W})^\top (\boldsymbol{I}_d \otimes \boldsymbol{C}^{(t)}) \mathrm{vec}(\boldsymbol{W}), \end{aligned}$$

where $\boldsymbol{I}_d \in \mathbb{R}^{d\times d}$ *is the identity matrix,* $\mathrm{tr}(\cdot)$ *is the trace operator, and* $\mathrm{vec}(\cdot)$ *is the vectorization operator.*

Lemma 4 *Under the updating rule of Eq.* (6)*,*

$$\widetilde{J}(\boldsymbol{W}^{(t+1)}) - \widetilde{J}(\boldsymbol{W}^{(t)}) \leq 0.$$

Proof: Under the updating rule of Eq. (6)*, since Eq.*(5) *is a convex function and the optimal solution is obtained by solving* $\frac{\partial \widetilde{J}(\boldsymbol{W})}{\partial \boldsymbol{W})} = 0$*, the obtained solution* $\boldsymbol{W}^{(t+1)}$ *is the global solution. That is,* $\widetilde{J}(\boldsymbol{W}^{(t+1)}) \leq \widetilde{J}(\boldsymbol{W}^{(t)})$*.*

Lemma 5 *For any nonzero vectors $\boldsymbol{w}, \boldsymbol{w}^{(t)} \in \mathbb{R}^d$, the following inequality holds Nie et al. [2010]:*

$$\|\boldsymbol{w}\|_2 - \frac{\|\boldsymbol{w}\|_2^2}{2\|\boldsymbol{w}^{(t)}\|_2} \leq \|\boldsymbol{w}^{(t)}\|_2 - \frac{\|\boldsymbol{w}^{(t)}\|_2^2}{2\|\boldsymbol{w}^{(t)}\|_2}.$$

Lemma 6 *For $r_{i,j} \geq 0, \forall i, j$, the following inequality holds for any non-zero vectors $\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}, \boldsymbol{w}_i^{(t+1)} - \boldsymbol{w}_j^{(t+1)}$:*

$$\begin{aligned}
&\sum_{i,j=1}^{n} r_{ij}\|\boldsymbol{w}_i^{(t+1)} - \boldsymbol{w}_j^{(t+1)}\|_2 - \mathrm{vec}(\boldsymbol{W}^{(t+1)})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t+1)}) \\
&- \left(\sum_{i,j=1}^{n} r_{ij}\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2 - \mathrm{vec}(\boldsymbol{W}^{(t)})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t)}) \right) \leq 0.
\end{aligned}$$

Proof: $\mathrm{vec}(\boldsymbol{W})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W})$ *can be written as*

$$\mathrm{vec}(\boldsymbol{W})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W}) = \sum_{i,j=1}^{n} r_{ij} \frac{\|\boldsymbol{w}_i - \boldsymbol{w}_j\|_2^2}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2}.$$

where $r_{ij} \geq 0$.

Then, the left hand side equation can be written as

$$\begin{aligned}
\Delta_g = &\sum_{i,j=1}^{n} r_{ij} \left(\|\boldsymbol{w}_i^{(t+1)} - \boldsymbol{w}_j^{(t+1)}\|_2 - \frac{\|\boldsymbol{w}_i^{(t+1)} - \boldsymbol{w}_j^{(t+1)}\|_2^2}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} \right) \\
&- \sum_{i,j=1}^{n} r_{ij} \left(\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2 - \frac{\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2^2}{2\|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2} \right).
\end{aligned}$$

Using Lemma 5, $\Delta_g \leq 0$. □

Lemma 7 *The following inequality holds for any non-zero vectors Kong et al. [2014]:*

$$\begin{aligned}
&\sum_{i=1}^{n} \|\boldsymbol{w}_i^{(t+1)}\|_1^2 - \mathrm{vec}(\boldsymbol{W}^{(t+1)})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t+1)}) \\
&- \left(\sum_{i=1}^{n} \|\boldsymbol{w}_i^{(t)}\|_1^2 - \mathrm{vec}(\boldsymbol{W}^{(t)})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t)}) \right) \leq 0.
\end{aligned} \tag{1}$$

Proof: $\mathrm{vec}(\boldsymbol{W})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W})$ *can be written as*

$$\begin{aligned}
\mathrm{vec}(\boldsymbol{W})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W}) &= \sum_{\ell=1}^{dn} [\mathrm{vec}(\boldsymbol{W})]_\ell^2 \sum_{i=1}^{n} \frac{I_{i,\ell}\|\boldsymbol{w}_i^{(t)}\|_1}{[\mathrm{vec}(|\boldsymbol{W}^{(t)}|)]_\ell} \\
&= \sum_{i=1}^{n} \left(\sum_{j=1}^{d} \frac{[\boldsymbol{w}_i]_j^2}{[|\boldsymbol{w}_i^{(t)}|]_j} \right) \|\boldsymbol{w}_i^{(t)}\|_1.
\end{aligned}$$

Thus, the left hand equation is written as

$$\begin{aligned}
\Delta_e &= \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{d} [\boldsymbol{w}_i^{(t+1)}]_j \right)^2 - \left(\sum_{j=1}^{d} \frac{[\boldsymbol{w}_i^{(t+1)}]_j^2}{[|\boldsymbol{w}_i^{(t)}|]_j} \right) \left(\sum_{j=1}^{d} [\boldsymbol{w}_i^{(t)}]_j \right) \right] \\
&= \sum_{i=1}^{n} \left[\left(\sum_{j=1}^{d} a_j^{(t)} b_j^{(t)} \right)^2 - \left(\sum_{j=1}^{d} (a_j^{(t)})^2 \right) \left(\sum_{j=1}^{d} (b_j^{(t)})^2 \right) \right] \leq 0,
\end{aligned}$$

where $a_j^{(t)} = \frac{[|\boldsymbol{w}_i^{(t+1)}|]_j}{\sqrt{[|\boldsymbol{w}_i^{(t)}|]_j}}$ *and* $b_j^{(t)} = \sqrt{[|\boldsymbol{w}_i^{(t)}|]_j}$, *and* $\mathrm{vec}(\boldsymbol{W}^{(t)})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t)}) = \sum_{i=1}^n \|\boldsymbol{w}_i^{(t)}\|_1^2$. *The inequality holds due to cauchy inequality Steele [2004].* □

Lemma 8 *For* $r_{i,j} \geq 0, \forall i, j$, *the following inequality holds for any non-zero vectors* $\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}, \boldsymbol{w}_i^{(t+1)} - \boldsymbol{w}_j^{(t+1)}$:

$$J(\boldsymbol{W}^{(t+1)}) - J(\boldsymbol{W}^{(t)}) \leq \widetilde{J}(\boldsymbol{W}^{(t+1)}) - \widetilde{J}(\boldsymbol{W}^{(t)}).$$

Proof: The difference between the right and left side equations is given as

$$\begin{aligned}
\Delta &= J(\boldsymbol{W}^{(t+1)}) - J(\boldsymbol{W}^{(t)}) - (\widetilde{J}(\boldsymbol{W}^{(t+1)}) - \widetilde{J}(\boldsymbol{W}^{(t)})) \\
&= \lambda_1 \Big(\sum_{i,j=1}^n r_{ij} \|\boldsymbol{w}_i^{(t+1)} - \boldsymbol{w}_j^{(t+1)}\|_2 - \mathrm{vec}(\boldsymbol{W}^{(t+1)})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t+1)}) \\
&\quad - \big[\sum_{i,j=1}^n r_{ij} \|\boldsymbol{w}_i^{(t)} - \boldsymbol{w}_j^{(t)}\|_2 - \mathrm{vec}(\boldsymbol{W}^{(t)})^\top \boldsymbol{F}_g^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t)}) \big] \Big) \\
&\quad + \lambda_2 \Big(\sum_{i=1}^n \|\boldsymbol{w}_i^{(t+1)}\|_1^2 - \mathrm{vec}(\boldsymbol{W}^{(t+1)})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t+1)}) \\
&\quad - \big[\sum_{i=1}^n \|\boldsymbol{w}_i^{(t)}\|_1^2 - \mathrm{vec}(\boldsymbol{W}^{(t)})^\top \boldsymbol{F}_e^{(t)} \mathrm{vec}(\boldsymbol{W}^{(t)}) \big] \Big)
\end{aligned}$$

Based on Lemma 6 and 7, $\Delta \leq 0$. □

References

Deguang Kong, Ryohei Fujimaki, Ji Liu, Feiping Nie, and Chris Ding. Exclusive feature learning on arbitrary structures via ℓ_{12}-norm. In *NIPS*, 2014.

Feiping Nie, Heng Huang, Xiao Cai, and Chris H Ding. Efficient and robust feature selection via joint $\ell_{2,1}$-norms minimization. In *NIPS*, 2010.

J Michael Steele. An introduction to the art of mathematical inequalities: The Cauchy-Schwarz master class, 2004.