Convex relaxations of sparsity

\[ \|w\|_1^0 = \min \left\{ \sum_{i=1}^{L} |r_i| : w = \sum_{i=1}^{L} r_i \text{ Support}(r_i) \right\}, \]
where \( \text{Support}(r_i) \) is (all subsets of \( \{1, \ldots, d\} \) of size \( k \)).

This norm is the tightest convex relaxation of \([w : \|w\|_0 \leq k, \|w\|_1 \leq 1]\).

Computing the norm in \( O(d|\log(d)|) \) steps:
for \( w \geq \cdots \geq w_1 \geq 0 \), first find the unique \( r \in \{0, \ldots, k-1\} \) s.t.
\[ w_{k-r+1} > \frac{\beta + 1}{\beta - k + (\beta + 1)(r + 1)} \sum_{j=k-r}^{l} w_j \geq w_{k-r}, \]
Then
\[ \|w\|_1^0 = \sum_{i=1}^{r-1} \|w_i\|_2^2 + \sum_{i=r}^{l} \|w_i\|_1 \]

\( f_j \) norm on larger terms
\( f_t \) norm on smaller terms

Equivalent to an overlapping group lasso [4] where the set of groups is \( \{\text{all groups of size up to } k\} \), but computing it this way would take exponential time.

The dual norm is the \( f_j \) norm of the \( k \) largest terms:
\[ \|w\|_1^0 = \max \{\|w_i\|_2 / l : i \in \mathbb{G}\}. \]

Optimization

- \( f(w) = L \)-Lipschitz, continuously-differentiable loss
- Want to minimize \( f(w) + \lambda \cdot (\|w\|_0^0)^2 \)

For each step of the FISTA algorithm [4], we need to be able to compute the proximity operator

Algorithm:
WLOG \( w_j \geq \cdots \geq w_1 \geq 0 \), find \( r \in \{0, \ldots, k-1\} \) s.t.
\[ w_{k-r+1} > \frac{\beta + 1}{\beta - k + (\beta + 1)(r + 1)} \sum_{j=k-r}^{l} w_j \geq w_{k-r}, \]
Then set
\[ z_j = \frac{\bar{w}_j}{\lambda}, \]
for \( j = 1, \ldots, r-1 \) (relative shrinkage)

\[ z_j = 1 - \frac{\bar{w}_j}{\lambda}, \]
for \( j = k, \ldots, d \) (soft thresholding)

Cone of \( k \)-support norm

The \( k \)-support norm

- \( f(w) = L \)-Lipschitz, continuously-differentiable loss
- Want to minimize \( f(w) + \lambda \cdot (\|w\|_0^0)^2 \)

For each step of the FISTA algorithm [4], we need to be able to compute the proximity operator

Algorithm:
WLOG \( w_j \geq \cdots \geq w_1 \geq 0 \), find \( r \in \{0, \ldots, k-1\} \) s.t.
\[ w_{k-r+1} > \frac{\beta + 1}{\beta - k + (\beta + 1)(r + 1)} \sum_{j=k-r}^{l} w_j \geq w_{k-r}, \]
Then set
\[ z_j = \frac{\bar{w}_j}{\lambda}, \]
for \( j = 1, \ldots, r-1 \) (relative shrinkage)

\[ z_j = 1 - \frac{\bar{w}_j}{\lambda}, \]
for \( j = k, \ldots, d \) (soft thresholding)

Cone of \( k \)-support norm

The \( k \)-support norm

- \( f(w) = L \)-Lipschitz, continuously-differentiable loss
- Want to minimize \( f(w) + \lambda \cdot (\|w\|_0^0)^2 \)

For each step of the FISTA algorithm [4], we need to be able to compute the proximity operator

Algorithm:
WLOG \( w_j \geq \cdots \geq w_1 \geq 0 \), find \( r \in \{0, \ldots, k-1\} \) s.t.
\[ w_{k-r+1} > \frac{\beta + 1}{\beta - k + (\beta + 1)(r + 1)} \sum_{j=k-r}^{l} w_j \geq w_{k-r}, \]
Then set
\[ z_j = \frac{\bar{w}_j}{\lambda}, \]
for \( j = 1, \ldots, r-1 \) (relative shrinkage)

\[ z_j = 1 - \frac{\bar{w}_j}{\lambda}, \]
for \( j = k, \ldots, d \) (soft thresholding)