# STRONG RULES FOR EFFICIENT LASSO COMPUTATIONS AND THE BASIL ALGORITHM

NOTES BY PARKER KNIGHT

## 1. Strong rules for the LASSO

1.1. **Subgradients.** Let $f : \mathbb{R}^m \to \mathbb{R}$ be convex.

Recall the following first order condition: if $f$ is differentiable, then

$$f(y) \geq f(x) + \nabla f(x)^T (y - x) \quad \forall x, y \in \text{dom}(f)$$

What if $f$ is not differentiable? This motivates the following definition: Call $g \in \mathbb{R}^m$ a *subgradient* of $f$ at $x$ iff

$$f(y) \geq f(x) + g^T (y - x) \quad \forall x, y \in \text{dom}(f)$$

The subdifferential of $f$ at $x$, denote $\partial f(x)$, is the set of all subgradients. The following facts will be useful:

(1) If $f$ is differentiable at $x$, then $\partial f(x) = \{\nabla f(x)\}$
(2) For $\alpha_1, \alpha_2 \geq 0$, then $\partial \left[\partial \alpha_1 f_1(x) + \alpha_2 f_2(x)\right] = \alpha_1 \partial f_1(x) + \alpha_2 \partial f_2(x)$
(3) $x^*$ minimizes $f$ iff $0 \in \partial f(x^*)$

where we define set addition as $A + B = \{a + b | a \in A, b \in B\}$.

1.2. **The LASSO.** Recall the LASSO loss function:

$$Q_\lambda(\beta) = \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

where the LASSO solution $\hat{\beta}(\lambda)$ satisfies

$$Q_\lambda(\hat{\beta}(\lambda)) = \min_\beta Q_\lambda(\beta)$$

Note that $Q_\lambda(.)$ is not differentiable everywhere, but it is convex. So $\hat{\beta}_\lambda$ minimizes $Q$ iff $0 \in \partial Q_\lambda(\hat{\beta}(\lambda))$.

But how do we find $\partial Q$?

$$\partial Q_\lambda(b) = -X^T (Y - Xb) + \lambda \gamma$$

where $\gamma \in \partial ||b||_1$. For $g(x) = |x|$, we have

$$\partial g(x) = \begin{cases} \{1\} & x > 0 \\ \{-1\} & x < 0 \\ [-1, 1] & x = 0 \end{cases}$$

since $g$ is differentiable when $x \neq 0$, and when $x = 0$ we have $|y| \geq \alpha y$ iff $\alpha \in [-1, 1]$. We extend this component-wise[1] to get the expression for $\gamma$:

$$\gamma_j \in \begin{cases} \{1\} & b_j > 0 \\ \{-1\} & b_j < 0 \\ [-1, 1] & b_j = 0 \end{cases}$$

for $j = 1, ..., p$. Using this, we have that $\hat{\beta}(\lambda)$ minimizes $Q$ iff

$$X_j^T(Y - X\hat{\beta}(\lambda)) = \lambda\gamma_j$$

for $j = 1, ..., p$ with $\gamma_j$ defined as above for $\hat{\beta}(\lambda)$. This admits a key detail: $|X_j^T(Y - X\hat{\beta}(\lambda))| < \lambda$ implies $\hat{\beta}_j(\lambda) = 0$.

### 1.3. **Strong rules.**

## 2. BASIL

## References

[1] Proof is simple, but requires a bit more subgradient calculus to do formally.