

Subgradient Methods Applied to LASSO Regression

Kasey Tian *Electrical and Computer Engineering*
Rutgers University
 New Brunswick, NJ, USA
 kasey.tian@rutgers.edu

Abstract

Subgradient methods

Index Terms

optimization, subgradients, LASSO, regression

I. INTRODUCTION

Subgradient methods are a way to perform an optimization on an objective function which is not fully differentiable. For non-differentiable objective functions, traditional methods, such as gradient descent and Newton's method, are impossible to execute. They can also be combined with a wide variety of other optimization methods, and have far reaching applications. They were originally developed by Shor and others in the Soviet Union in the 1960s and 70s [1] [2]. LASSO regression is a regression analysis method which was first introduced in 1986 [3] in the field of geophysics, but was then independently rediscovered, named, and popularized in 1996 by Tibshirani [4]. It is characterized by regularization using the L_1 norm, which involves the absolute value function and is therefore not fully differentiable. This makes it a prime candidate for applying subgradient methods. To demonstrate a practical implementation LASSO regression will be used with a linear classifier to predict life expectancy from World Health Organization (WHO) data.

II. MATHEMATICAL ANALYSIS

In this section we will explore the mathematics of subgradients, subgradient methods and LASSO regression.

A. Defining Subgradient

This section is mostly adapted from [5], with some supplementary material from [1].

1) *Definitions:* A subgradient is defined for some convex function $f : \mathbb{R}^n \rightarrow \mathbb{R}$ at a point $x \in \text{dom } f$ as a vector $g \in \mathbb{R}^n$ such that $\forall y \in \text{dom } f$

$$f(y) \geq f(x) + g^T(y - x) \quad (1)$$

Alternately expressed as

$$f(y) - f(x) \geq g^T(y - x) \quad (2)$$

$$f(x) - f(y) + g^T(y - x) \leq 0 \quad (3)$$

There can be multiple subgradients at a point x , so we will also define the subdifferential $\partial f(x)$ to be the set of all subgradients at x .

$$\partial f(x) = \bigcap_{y \in \text{dom } f} \{g : f(y) \geq f(x) + g^T(y - x)\} \quad (4)$$

If there exists at least one subgradient at a point x , we would say f is subdifferentiable at x . If all points in the domain are subdifferentiable, we say that f is subdifferentiable.

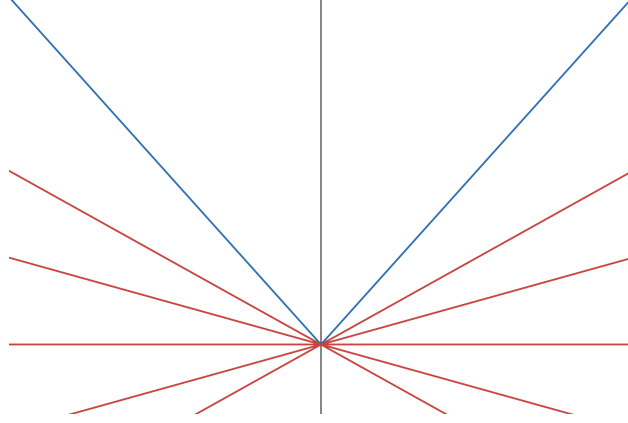


Fig. 1. Absolute value function (blue) with subgradients (red)

2) *Example: Absolute Value:* If we consider g in (1) to be a slope, we can visualize a subgradient as being some hyperplane intersecting our function at x for which all values of the function are on or above the plane.

Let us employ this intuition to find a subgradient of the function $f(x) = |x|$ at the point $x = 0$. Graphically, we can see in Fig. 1 that many different lines satisfy this criterion. In fact, we can say that any $g \in [-1, 1]$ would be a subgradient, and therefore $\partial f(0) = [-1, 1]$. But what about other points? For a point $x > 0$, we can surmise that the only possible $g = 1$, as any other value will leave some parts of our function beneath the resulting plane. Likewise for $x < 0$, $g = -1$. Using (4), we can say

$$\partial f(x) = \begin{cases} \{-1\} & x < 0 \\ [-1, 1] & x = 0 \\ \{1\} & x > 0 \end{cases} \quad (5)$$

This can be compared against the derivative of $f(x)$.

$$f'(x) = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \end{cases} \quad (6)$$

We find that where the function is differentiable, the subdifferential contains only the gradient. There is only a difference where the function is not differentiable. Here we find that our set ranges between the two derivatives on either side of it. More formally, we can say that $\partial f(x) = [a, b]$ such that

$$a = \lim_{y \rightarrow x^-} \frac{f(y) - f(x)}{y - x} \quad (7)$$

$$b = \lim_{y \rightarrow x^+} \frac{f(y) - f(x)}{y - x} \quad (8)$$

for one dimensional functions. We can note that for differentiable points $a = b$, so this definition satisfies our observation up to this point.

3) *Properties:* There are a few important properties that we shall take note of. Proofs of these properties can be found in Appendix A.

- If and only if the function is differentiable at x , $\partial f(x) = \{\nabla f(x)\}$
- If and only if x^* is a global minimum, $0 \in \partial f(x^*)$
- $\partial f(x)$ will always be a convex set

B. Subgradient Methods

This section is predominantly adapted from [1] and [2].

1) *Iteration Step:* Subgradient methods are a family of optimization techniques involving the same basic iteration step

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} \quad (9)$$

where $g^{(k)}$ is any subgradient at the point $x^{(k)}$ and $\alpha_k > 0$. There are a number of different methods of choosing α_k , which will be further explored in Section II-C. This formula can be seen to be very similar to gradient descent. In fact, can be observed that since the only subgradient at differentiable points is the gradient, all subgradient methods are gradient descent in these sections, the only difference from regular gradient descent being our way of choosing step size. One notable difference from gradient descent is that subgradient methods are not descent methods; we are not guaranteed that every step will descend. For that reason we should have some method of tracking our best performing point.

$$f_{best}^{(k)} = \min(f_{best}^{(k-1)}, f(x^{(k)})) \quad (10)$$

And a corresponding $i_{best}^{(k)}$ such that $f(x^{(i_{best}^{(k)})}) = f_{best}^{(k)}$. This would not be needed in a descent method, because our best performing point is always the last point.

2) *Convergence Proof:* We are focused on converging towards the optimum point within some error of margin. We begin by making the following assumptions

- There exists a minimizer of f called x^* , for which $f^* = f(x^*)$
- There exists some bounding value G for which $\|g\|_2 \leq G, g \in \partial f(x), \forall x \in \text{dom } f$ ¹²
- There exists some known bounding value R such that $R \geq \|x^{(1)} - x^*\|_2$

Next we can write the following based on (9)

$$\|x^{(k+1)} - x^*\|_2^2 = \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \quad (11)$$

Focusing on the right side, we can multiply the quadratic out

$$\|x^{(k+1)} - x^*\|_2^2 = \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \quad (12)$$

Using the definition of subgradient from (2), we can write

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \quad (13)$$

Paying particular attention to the term $\|x^{(k)} - x^*\|_2^2$, we can notice we have created a recursive inequality. If we perform a recursive substitution until reaching $x^{(1)}$ we would find ourselves with the following inequality

$$\|x^{(k+1)} - x^*\|_2^2 \leq \|x^{(1)} - x^*\|_2^2 - 2 \sum_{i=1}^k \alpha_i (f(x^{(i)}) - f^*) + \sum_{i=1}^k \alpha_k^2 i \|g^{(i)}\|_2^2 \quad (14)$$

We know that because of our bounding value $R \geq \|x^{(1)} - x^*\|_2$ and because of the norm, $\|x^{(k+1)} - x^*\|_2^2 \geq 0$, so we can write

$$2 \sum_{i=1}^k \alpha_i (f(x^{(i)}) - f^*) \leq R^2 + \sum_{i=1}^k \alpha_k^2 i \|g^{(i)}\|_2^2 \quad (15)$$

It follows readily from (10) that $f_{best}^{(k)} \leq f(x^{(i)}) \forall 1 \leq i \leq k$, so we can say

$$2 \sum_{i=1}^k \alpha_i (f_{best}^{(k)} - f^*) \leq 2 \sum_{i=1}^k \alpha_i (f(x^{(i)}) - f^*) \leq R^2 + \sum_{i=1}^k \alpha_k^2 i \|g^{(i)}\|_2^2 \quad (16)$$

And therefore

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + \sum_{i=1}^k \alpha_k^2 i \|g^{(i)}\|_2^2}{2 \sum_{i=1}^k \alpha_i} \quad (17)$$

We can finish by applying our G bounding condition to $\|g^{(k)}\|_2$, letting us write the final inequality

$$f_{best}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^k \alpha_k^2 i}{2 \sum_{i=1}^k \alpha_i} \quad (18)$$

¹One way that this assumption can be met is if the function satisfies the Lipschitz condition, but it is not the only way

²This assumption helps us greatly in our analysis, but it is not strictly necessary. There exist subgradient methods which can be proven to work even when this assumption does not hold [1]

Given that R , G , and α_k are finite values, we can surmise that $f_{best}^{(k)}$ approaches f^* within some small error bound that is dependent on α_k and k . It is evident from this that the means we use to choose α_k will have significantly affect the error bound.

C. Choosing Step Size

There are many different methods of choosing the step size $\alpha^{(k)}$. We will be discussing five broad categories in this section.

The first is a constant step size, where for some constant $\alpha > 0$

$$\alpha^{(k)} = \alpha \quad (19)$$

Second is a constant step length, where for some constant $\gamma > 0$

$$\alpha^{(k)} = \frac{\gamma}{\|g^{(k)}\|_2} \quad (20)$$

It can be easily proven that for this step size, $\|x^{(k+1)} - x^{(k)}\|_2 = \gamma$

D. LASSO Regression

III. CODE IMPLEMENTATION

IV. NUMERICAL RESULTS

V. CONCLUSION

APPENDIX

A. Proving Subgradient Properties

1) *When the Function is Differentiable:* We can begin by supposing that f is differentiable at x , and therefore $\nabla f(x)$ exists. Let us write a definition for each of its elements

$$\nabla f_i(x) = \lim_{y \rightarrow x} \frac{f(y) - f(x)}{y_i - x_i} \quad (21)$$

And let us also rewrite (3) with an elementwise sum

$$f(x) - f(y) + \sum_{i=1}^n g_i^T (y_i - x_i) \leq 0 \quad (22)$$

We can combine (21) and (22)

$$\lim_{y \rightarrow x} f(x) - f(y) + \sum_{i=1}^n \nabla f_i(x)^T (y_i - x_i) \leq 0 \quad (23)$$

Substitute and simplify and we find that

$$\lim_{y \rightarrow x} f(x) - f(y) + \sum_{i=1}^n \nabla f_i(x)^T (y_i - x_i) = 0 \quad (24)$$

If we are interested in trying an alternate $g \neq \nabla f_i(x)$, we can say that a particular g_i would be invalid if

$$g_i(y_i - x_i) > \nabla f_i(x)(y_i - x_i) \quad (25)$$

If $g_i > \nabla f_i(x)$ and $y_i > x_i$, this condition is fulfilled and therefore no values of $g_i > \nabla f_i(x)$ are valid. Likewise, if $g_i < \nabla f_i(x)$ and $y_i < x_i$ this condition is fulfilled and therefore no values of $g_i < \nabla f_i(x)$ are valid. Therefore the only valid values for every $g_i = \nabla f_i(x)$, so the only valid $g = \nabla f(x)$ if $\nabla f(x)$ exists.

2) *Global Minimum:* If x^* is a global minimum it must be true that $\forall y \in \text{dom } f$

$$f(x^*) \leq f(y) \quad (26)$$

If we have g equal to the zero vector in (1), we end up removing the second term, so we end up with

$$f(y) \geq f(x) \quad (27)$$

We can trivially redefine $x = x^*$ to derive (26). Therefore x^* is optimal if and only if the zero vector is included in its subdifferential.

3) *Convexity*: We can view (3) as defining a halfspace in terms of x, y, f . From (4) we can see that $\partial f(x)$ is composed of intersections of these halfspaces. Since halfspaces are always convex and the intersection of convex sets is always convex, $\partial f(x)$ must be a convex set.

REFERENCES

- [1] S. Boyd and J. Park, "Subgradient methods," May 2014. [Online]. Available: https://web.stanford.edu/class/ee364b/lectures/subgrad_method_notes.pdf
- [2] S. Boyd, L. Xiao, and A. Mutapcic, "Subgradient methods," October 2003. [Online]. Available: https://web.mit.edu/6.976/www/notes/subgrad_method.pdf
- [3] F. Santosa and W. W. Symes, "Linear inversion of band-limited reflection seismograms," *SIAM Journal on Scientific and Statistical Computing*, vol. 7, no. 4, pp. 1307–1330, 1986. [Online]. Available: <https://doi.org/10.1137/0907087>
- [4] R. Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society: Series B (Methodological)*, vol. 58, no. 1, pp. 267–288, 12 2018. [Online]. Available: <https://doi.org/10.1111/j.2517-6161.1996.tb02080.x>
- [5] S. Boyd and L. Vandenberghe, "Subgradients," April 2008. [Online]. Available: https://see.stanford.edu/materials/lsocoe364b/01-subgradients_notes.pdf