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.

II. MATHEMATICAL BASIS

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

A. Subgradient

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

$$f(y) \geq f(x) + g^T(y - x) \tag{1}$$

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.

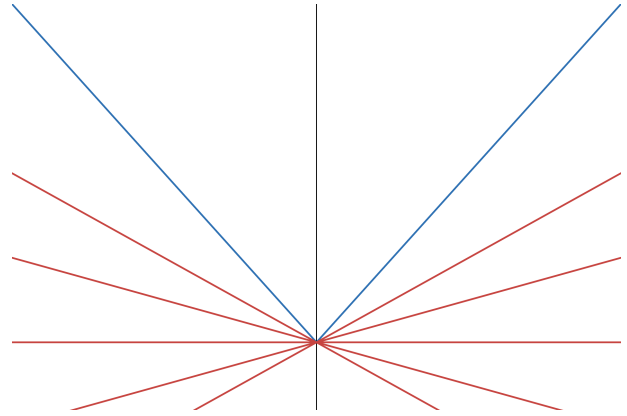$$\partial f(x) = \left\{ g : f(y) \geq f(x) + g^T(y - x) \right\} \tag{2}$$

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 criteron. 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 $x < 0$, $g = -1$. Thus we can write

$$\partial f(x) = \begin{cases} -1 & x < 0 \\ [-1, 1] & x = 0 \\ 1 & x > 0 \end{cases} \tag{3}$$

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

$$f'(x) = \begin{cases} -1 & x < 0 \\ 1 & x > 0 \end{cases} \tag{4}$$

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

$$a = \lim_{y \to x^-} \frac{f(y) - f(x)}{y - x} \qquad (5)$$

$$b = \lim_{y \to x^+} \frac{f(y) - f(x)}{y - x} \qquad (6)$$

for one dimensional functions. We can note that this set is convex, which will hold true for all subdifferentials. Furthermore, we can find the boundaries of this set using limits neae our non-differentiable points.

3) Properties: There are a few important properties that we shall take note of.

1) If the function is differentiable at x, the only member of $\partial f(x)$ will be the gradient
2) If x^* is a global minimum, $\partial f(x^*)$ must contain the zero vector
3) $\partial f(x)$ will always be a convex set

B. Subgradient Methods

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

$$x^{(k+1)} = x^{(k)} - \alpha^{(k)} g^{(k)} \qquad (7)$$

where $g^{(k)}$ is any subgradient at the point $x^{(k)}$. 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.

$$f_{best}^{(k)} = \min(f_{best}^{(k-1)}, f(x^{(k)})) \qquad (8)$$

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) Choosing a Step Size: There are a few different methods of choosing the step size $\alpha^{(k)}$

1) Constant step size: $\alpha^{(k)} = \alpha$ for a fixed α, regardless of the value of k.

C. LASSO Regression

III. Code Implementation

IV. Numerical Results

V. Conclusion

Appendix

A. Subgradient Properties

1) :

B. Second Appendix

ha ha ha

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."

[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