Subgradient Methods on LASSO Regression

Kasey Tian

16:332:509

May 5, 2025

Outline

Defining Subgradient

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

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

Defining Subdifferential

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.

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

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. [1]

Example: Absolute Value

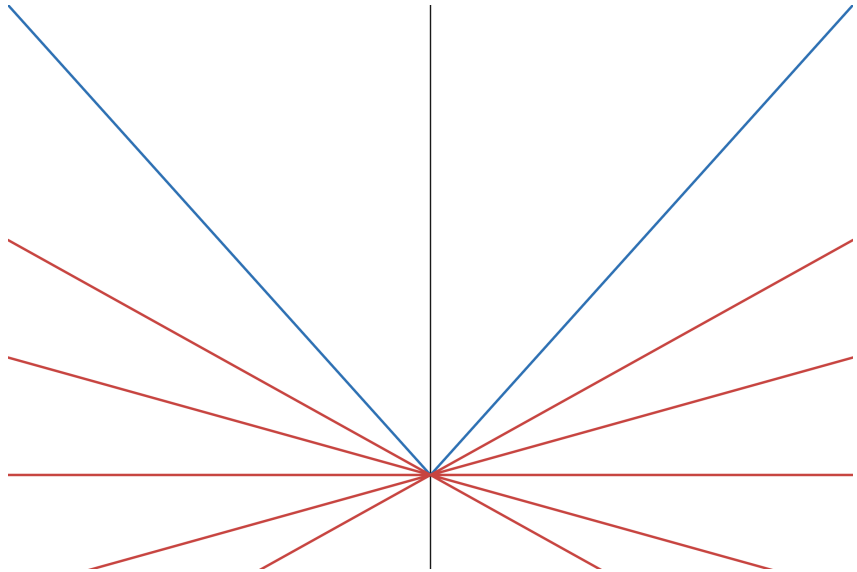

Figure: Absolute value function (blue) with subgradients (red)

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

Subgradient Method Iteration

Subgradient methods are a family of optimization techniques. They were invented in the 1960s and 70s in the Soviet Union. They all involve the same basic iteration step [2]

$$x^{(k+1)} = x^{(k)} - \alpha_k g^{(k)} \tag{4}$$

where $g^{(k)}$ is any subgradient at the point $x^{(k)}$ and $\alpha_k > 0$

Subgradient Method Properties

Subgradient methods are unlike gradient descent and Newton's method in that it is not a descent method. We are not guaranteed that an iteration will descend. Therefore we need to keep track of the best solution we have so far.

$$f_{\text{best}}^{(k)} = \min(f_{\text{best}}^{(k-1)}, f(x^{(k)})) \tag{5}$$

And a corresponding $i_{\text{best}}^{(k)}$ such that $f(x^{(i_{\text{best}}^{(k)})}) = f_{\text{best}}^{(k)}$

Subgradient Method Convergence Proof - Assumptions

There exists a minimizer of f called x^*, for which $f^* = f(x^*)$

There exists some bounding value G for which $\|g\|_2 \leq G, \forall\, g \in \partial f(x), \forall\, x \in \mathbf{dom}\, f$[1][2]

There exists some known bounding value R such that $R \geq \|x^{(1)} - x^*\|_2$

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

[2] This 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 [2]

Subgradient Method Convergence Proof - Step 1

$$
\begin{aligned}
\|x^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - \alpha_k g^{(k)} - x^*\|_2^2 \\
\|x^{(k+1)} - x^*\|_2^2 &= \|x^{(k)} - x^*\|_2^2 - 2\alpha_k g^{(k)T}(x^{(k)} - x^*) + \alpha_k^2 \|g^{(k)}\|_2^2 \\
\|x^{(k+1)} - x^*\|_2^2 &\leq \|x^{(k)} - x^*\|_2^2 - 2\alpha_k (f(x^{(k)}) - f^*) + \alpha_k^2 \|g^{(k)}\|_2^2
\end{aligned} \tag{6}
$$

Subgradient Method Convergence Proof - Step 2

$$\|x^{(k+1)} - x^*\|_2^2 \leq$$

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

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

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

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

$$f_{\text{best}}^{(k)} - f^* \leq \frac{R^2 + G^2 \sum_{i=1}^{k} \alpha_i^2}{2\sum_{i=1}^{k} \alpha_i} \tag{7}$$

How to Choose Step Size

The step size can be chosen in many different ways, we will focus on 5 categories

- Constant Step Size
- Constant Step Length
- Square Summable But Not Summable
- Nonsummable Diminishing Step Size
- Nonsummable Diminishing Step Length

Step Sizes- Constant

Constant Step Size:

$$\alpha_k = \alpha$$
$$\lim_{k\to\infty} \frac{R^2 + G^2\alpha^2 k}{2\alpha k} = \frac{G^2\alpha}{2} \tag{8}$$

Constant Step Length:

$$\alpha_k = \frac{\gamma}{\|g^{(k)}\|_2}$$
$$\lim_{k\to\infty} \frac{R^2 + \gamma^2 k}{2\gamma k/G} = \frac{G\gamma}{2} \tag{9}$$

Step Sizes - Nonsummable

Square Summable Not Summable:

$$\alpha_k \geq 0 \quad \|\alpha\|_2^2 = \sum_{k=1}^{\infty} \alpha_k^2 < \infty \quad \sum_{k=1}^{\infty} \alpha_k = \infty \tag{10}$$

Nonsummable Diminishing Step Size:

$$\alpha_k \geq 0 \quad \lim_{k \to \infty} \alpha_k = 0 \quad \sum_{k=1}^{\infty} \alpha_k = \infty \tag{11}$$

Nonsummable Diminishing Step Length:

$$\alpha_k = \frac{\gamma_k}{\|g^{(k)}\|_2} \quad \gamma_k \geq 0 \quad \lim_{k \to \infty} \gamma_k = 0 \quad \sum_{k=1}^{\infty} \gamma_k = \infty \tag{12}$$

LASSO Regression History

LASSO is short for **L**east **A**bsolute **S**hrinkage and **S**election **O**perator. It was technically first discovered in 1986 by geophysicists, but was later rediscovered, named, and popularized in 1996. It was originally formulated for use with linear classification models, but can be applied to other classifiers as well [3] [4]

Problem

The problem is p dimensional with a scalar outcome

We analyze n cases at a time

Collected inputs form the matrix $X \in \mathbb{R}^{n \times p}$

Collected outputs form the vector $y \in \mathbb{R}^n$

Parameters are $\alpha \in \mathbb{R}$, $\beta \in \mathbb{R}^p$

Our classifier takes the form $y = \beta^T x + \alpha$

LASSO Estimate

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin} \left\{ \sum_{i=1}^{n} \left(y_i - \alpha - \sum_{j=1}^{p} \beta_j X_{i,j} \right)^2 \right\} \tag{13}$$
$$\text{s.t.} \sum_{j=1}^{p} |\beta_j| \leq t$$

$$(\hat{\alpha}, \hat{\beta}) = \text{argmin} \, \|y - \alpha - X\beta\|_2^2 \tag{14}$$
$$\text{s.t.} \|\beta\|_1 - t \leq 0$$

We can add some additional definitions and assumptions. We assume that X is standardized and we have a vector $\bar{x}$ and a value $\bar{y}$ such that

$$\begin{aligned} \bar{x}_j = \sum_{i=1}^{n} X_{i,j}/n &= 0 \\ \sum_{i=1}^{n} X_{i,j}^2/n &= 1 \\ \bar{y} = \sum_{i=0}^{n} y_i/n &= 0 \end{aligned} \tag{15}$$

Estimates

$$\hat{\alpha} = 0$$
$$\hat{\beta} = \text{argmin} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \tag{16}$$

Subgradient of LASSO

$$
\begin{aligned}
\hat{\beta} &= \text{argmin} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \\
g &= \frac{\partial}{\partial \beta} \left\{ \frac{1}{n} \|y - X\beta\|_2^2 + \lambda \|\beta\|_1 \right\} \\
&= \frac{2}{n}(y - X\beta) \sum_{i=1}^{n} X_i + \text{sign}(\beta)\lambda
\end{aligned}
\tag{17}
$$

References I

[1] S. Boyd and L. Vandenberghe, *Subgradients*, Apr. 2008. [Online]. Available: https://see.stanford.edu/materials/lsocoee364b/01-subgradients_notes.pdf.

[2] S. Boyd and J. Park, *Subgradient methods*, May 2014. [Online]. Available: https://web.stanford.edu/class/ee364b/lectures/subgrad_method_notes.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. DOI: 10.1137/0907087. eprint: https://doi.org/10.1137/0907087. [Online]. Available: https://doi.org/10.1137/0907087.

References II

[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, Dec. 2018, ISSN: 0035-9246. DOI: 10.1111/j.2517-6161.1996.tb02080.x. eprint: https://academic.oup.com/jrsssb/article-pdf/58/1/267/49098631/jrsssb_58_1_267.pdf. [Online]. Available: https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.