EE 510: Mathematical Foundations of Machine Learning **Winter 2020**

Lecture Notes: Sparse Regression

Instructor Name: John Lipor

1 Introduction

Let's return to the matrix-vector version of least squares but keep the feature-label interpretation. In this case, we form $\mathbf{X}$ by letting the **rows** of $\mathbf{X}$ be the feature vectors $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^D$ corresponding to the labels $y_1, \ldots, y_N \in \mathbb{R}$. Recall that our goal is to regress $\mathbf{y} \in \mathbb{R}^N$ via linear combinations of the columns of $\mathbf{X}$, i.e., we want

$$\mathbf{X}\mathbf{w} \approx \mathbf{y},$$

where $\mathbf{w} \in \mathbb{R}^D$. In a large number of applications, we wish to represent $\mathbf{y}$ using the smallest number of columns of $\mathbf{X}$ possible. For example, if the columns of $\mathbf{X}$ are different features in a machine learning problem, then we may want to know the top few features that best predict our labels—a process known as **feature selection**.

1.1 Formulating sparse regression

We wish to encode the goal of minimizing the number of nonzero coefficients in $\mathbf{w}$ into a mathematical optimization problem (and ideally a convex one). Recall from the low-rank approximation slides that we defined the ℓ_0-"norm" of a vector $\|\mathbf{w}\|_0$ to be the number of nonzeros in that vector (norm is in quotes because it does not satisfy the definition of a **norm**). We can therefore encode our goal of sparse regression via the formulation

$$\min_{\mathbf{w} \in \mathbb{R}^D} \quad \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 \tag{1}$$

$$\text{subject to} \quad \|\mathbf{w}\|_0 \leq s \tag{2}$$

where $s \in \mathbb{N}$ is the number of nonzeros we allow in $\mathbf{w}$. Unfortunately, solving (2) is **NP-hard**, which is a precise way of saying that solving it is computationally prohibitive. This is in part due to the fact that (2) is not a convex problem. One way to see this is to note that for every $s \in \mathbb{N}$, there exists a $\lambda \in \mathbb{R}$ such that solving (2) is equivalent to solving

$$\min_{\mathbf{w} \in \mathbb{R}^D} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_0 . \tag{3}$$

Since $\|\cdot\|_0$ is not a convex function, the above is not a convex problem.

As we've learned, convex problems are much easier to deal with (typically through first-order methods). Our solution to the above problem is to find the convex problem whose solution approximates that of (3). Such a problem is called a **convex relaxation**. For reasons that you'll see in the homework, the natural convex relaxation of the ℓ_0-"norm" is the ℓ_1-norm (actually a norm), which is defined as

$$\|\mathbf{w}\|_1 = \sum_{i=1}^{D} |w_i| .$$

Using this, we form the convex relaxation of the problem (3) to be

$$\min_{\mathbf{w} \in \mathbb{R}^D} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{w}\|_1 , \tag{4}$$

which is called the **lasso** (least absolute shrinkage and selection operator). Since (4) attempts to regress $\mathbf{y}$ using a sparse vector $\mathbf{w}$, we refer to this problem as **sparse regression**. Further, we can interpret (4) in the context of empirical risk minimization, where the data-fit term (loss) is the usual least squares objective, and the regularizer is the ℓ_1-norm (instead of ℓ_2 as in ridge regression).

1.2 What leads to sparsity?

Why does the lasso lead to solutions that are sparse? Let's consider the constrained form of (4)

$$\min_{\mathbf{w}\in\mathbb{R}^D} \quad \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 \tag{5}$$

$$\text{subject to} \quad \|\mathbf{w}\|_1 \le s. \tag{6}$$

The constraint $\|\mathbf{w}\|_1 \le s$ means that the solution $\mathbf{w}^*$ lies in the ℓ_1-ball of radius s, i.e.,

$$\mathbf{w}^* \in \left\{\mathbf{w}\in\mathbb{R}^D : \|\mathbf{w}\|_1 \le s\right\}.$$

On the other hand, the set of solutions with equal regression error

$$\left\{\mathbf{w}\in\mathbb{R}^D : \|\mathbf{X}\mathbf{w}-\mathbf{y}\|_2^2 = c\right\}$$

is an ellipse for each $c\in\mathbb{R}$. The solution to (4) occurs for the smallest $c\in\mathbb{R}$ such that the elliptical contours hit the contraint set. For ℓ_1 regularization, these two sets tend to intersect at the corners where $\mathbf{w}$ is sparse, whereas ℓ_2 regularization (ridge regression) does not have this effect. For an excellent depiction, see **The Elements of Statistical Learning**, Fig. 3.11 (pg. 71).

1.3 Comparison to Ridge Regression

To build further intuition behind the lasso, let's consider the one-dimensional regression problem for standard least squares, ridge regression, and the lasso.

$$\begin{aligned}
\hat{w}_{LS} &= \arg\min_{w\in\mathbb{R}} (w-y)^2 \\
\hat{w}_{RR} &= \arg\min_{w\in\mathbb{R}} (w-y)^2 + \lambda w^2 \\
\hat{w}_{lasso} &= \arg\min_{w\in\mathbb{R}} (w-y)^2 + \lambda |w|.
\end{aligned}$$

When trying to optimize the above, one issue is that the absolute value is not differentiable. To overcome this, we introduce a new object.

Definition 1. *The **subdifferential** of a convex function $f:\mathbb{R}^D\to\mathbb{R}$ at $\mathbf{x}\in\mathbb{R}^D$ is*

$$\partial f(\mathbf{x}) = \left\{\mathbf{g}\in\mathbb{R}^D : f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \mathbf{g}, \mathbf{y}-\mathbf{x}\rangle \;\forall \mathbf{y}\in\mathbb{R}^D\right\}.$$

*A vector in $\mathbf{g}\in\partial f(\mathbf{x})$ is called a **subgradient**.*

Recall that for functions that are convex and differentiable, we had that

$$f(\mathbf{y}) \ge f(\mathbf{x}) + \langle \nabla f(\mathbf{x}), \mathbf{y}-\mathbf{x}\rangle$$

for all $\mathbf{y}\in\mathbb{R}^D$. Hence, we can see that for differentiable functions, the subdifferential consists of a single element, which is the gradient. Since $|w|$ is not differentiable at zero, we'll instead take a subgradient to optimize the one-dimensional lasso. You should check for yourself that a valid subgradient of $f(x)=|x|$ is

$g = 0$. With this subgradient in hand, we can write the closed-form solutions to our above three regression problems as

$$\begin{aligned}\hat{w}_{LS} &= y \\ \hat{w}_{RR} &= \frac{y}{1+\lambda} \\ \hat{w}_{lasso} &= \begin{cases} y-\lambda & y > \lambda \\ y+\lambda & y < -\lambda \\ 0 & y \in [-\lambda, \lambda] \end{cases}.\end{aligned}$$

These are obtained by setting the (sub)gradient to zero and solving for w and can be seen in Fig. 1 below. Both ridge regression and lasso are known as *shrinkage* methods, since they shrink the solution toward zero, with λ controlling the amount of shrinkage.

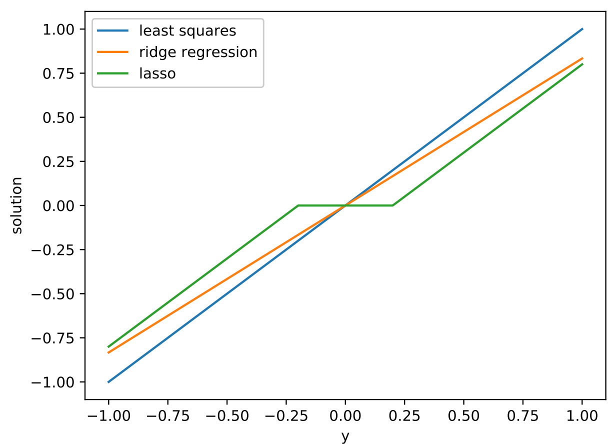

Figure 1: Solutions to the one-dimensional regression problem.