Lecture 23: Sparsity

Outline

What is sparsity?

Consider the following solution(s) to a linear regression problem:

$$w_1 = [0.1, 0.3, 0.5, 0.9, 0.1, 0.3, 0.4, 0.2]$$

What is sparsity?

Consider the following solution(s) to a linear regression problem:

$$w_1 = [0.1, 0.3, 0.5, 0.9, 0.1, 0.3, 0.4, 0.2]$$

$$w_2 = [0.0, 0.0, 0.5, 0.9, 0.0, 0.0, 0.4, 0.0]$$

What is sparsity?

Consider the following solution(s) to a linear regression problem:

$$w_1 = [0.1, 0.3, 0.5, 0.9, 0.1, 0.3, 0.4, 0.2]$$

$$w_2 = [0.0, 0.0, 0.5, 0.9, 0.0, 0.0, 0.4, 0.0]$$

We say w_2 is sparse, i.e., most of its weights are 0.

Why sparsity?

- Computational / memory efficiency (remove weights that are 0)
- Regularization / feature selection (zero weights correspond to features that are less important)

How to optimize for sparsity?

Start with least squares objective:

$$\min_{w} ||Xw - y||_2^2$$

How to optimize for sparsity?

Start with least squares objective:

$$\min_w ||Xw - y||_2^2$$

Add a constraint to enforce sparsity:

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

How to optimize for sparsity?

Start with least squares objective:

$$\min_{w} ||Xw - y||_2^2$$

Add a constraint to enforce sparsity:

$$\min_{w} ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

How to optimize for sparsity?

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

What's $||w||_0$?

How to optimize for sparsity?

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

What's $||w||_0$?

Hamming distance between w and $\vec{0}$.

I.e., the number of non-zero elements of w.

How to optimize for sparsity?

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

l_0 norm presents some challenges

How to optimize for sparsity?

$$\min_{w} ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

l_0 norm presents some challenges

1. It's not actually a norm (in particular, it's not convex)

How to optimize for sparsity?

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

l_0 norm presents some challenges

1. It's not actually a norm (in particular, it's not convex)
2. It's not differentiable

How to optimize for sparsity?

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

l_0 norm presents some challenges

1. It's not actually a norm (in particular, it's not convex)
2. It's not differentiable

--> Hard to use our standard optimization toolkit!

Outline

1. Motivation for sparsity
2. **The LASSO**
3. Matching pursuit
4. Orthogonal matching pursuit

The Least Absolute Shrinkage and Selection Operator (LASSO)

$$\min_{w} ||Xw - y||_2^2$$

The Least Absolute Shrinkage and Selection Operator (LASSO)

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_1 \leq k$$

The Least Absolute Shrinkage and Selection Operator (LASSO)

$$\min_{w} ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_1 \leq k$$

$$||w||_1 = \sum_{i=1}^{d} w_i$$

The Least Absolute Shrinkage and Selection Operator (LASSO)

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_1 \leq k$$

- Relaxation of the sparse objective (l_1 instead of l_2)
- Pros: l_1 is actually a norm
- Cons: Does this objective really induce sparsity?

Solving LASSO

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_1 \leq k$$

Step 1: convert LASSO problem into unconstrained form

Solving LASSO

$$\min_w ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_1 \leq k$$

Step 1: convert LASSO problem into unconstrained form

$$\min_w ||Xw - y||_2^2 + \lambda||w||_1$$

Solving LASSO

$$\min_{w} ||Xw - y||_2^2 + \lambda ||w||_1$$

Solving LASSO

$$\min_w ||Xw - y||_2^2 + \lambda ||w||_1$$

Observation: LASSO looks a lot like ridge regression!

$$\min_w ||Xw - y||_2^2 + \lambda ||w||_2$$

Solving LASSO

$$\min_w ||Xw - y||_2^2 + \lambda||w||_1$$

Can we just take gradients like in the ridge regression case?

Solving LASSO

$$\min_{w} ||Xw - y||_2^2 + \lambda||w||_1$$

Can we just take gradients like in the ridge regression case?

Not clear. $||w||_1$ is not differentiable.

Solving LASSO

$$\min_{w} ||Xw - y||_2^2 + \lambda ||w||_1$$

Solving LASSO

$$\min_w ||Xw - y||_2^2 + \lambda ||w||_1$$

- Objective is convex
- Still can optimize

Coordinate descent

Algorithm 1: Coordinate Descent

while $\mathbf{w}$ *has not converged* **do**

  pick a feature index i

  update w_i to $\arg\min_{w_i} L(\mathbf{w})$

Technical points:

- When does coordinate descent converge to the optimal solution?
- When L is **jointly convex** (and not if it's only **elementwise convex**)
- Fact: the LASSO objective is jointly convex

How to use coordinate descent to solve the LASSO problem?

Fact - the optimal solution has a closed form.

$r = \sum_{j \neq i} w_j x_j - y$

$$w_i^* = \begin{cases} 0 & \text{if } |a| \leq \lambda \\ \frac{-\lambda + a}{b} & \text{if } \frac{-\lambda+a}{b} > 0 \\ \frac{\lambda + a}{b} & \text{if } \frac{\lambda+a}{b} < 0 \end{cases}$$

$$a = -\sum_{j=1}^{n} 2x_{ji} r_j, \quad b = \sum_{j=1}^{n} 2x_{ji}^2$$

(To prove, take partial derivatives w.r.t. each coordinate considering the cases shown in the soln above

Relationship between LASSO and least squares solutions

Least squares solution:

$$w = (X^T X)^{-1} X^T y$$

Relationship between LASSO and least squares solutions

Least squares solution:

$$w = (X^T X)^{-1} X^T y$$

$$w_i = \frac{1}{\sum_{j=1}^{n} x_{ji}^2} x_{ji} \sum_{j \neq i} (y - w_j x_j)$$

Relationship between LASSO and least squares solutions

Least squares solution:

$$w = (X^T X)^{-1} X^T y$$

$$w_i = \frac{1}{\sum_{j=1}^{n} x_{ji}^2} x_{ji} \sum_{j \neq i} (y - w_j x_j)$$

$$w_i = \frac{a}{b}$$

--> Same as Lasso with $\lambda = 0$

Why does LASSO find a sparse solution?

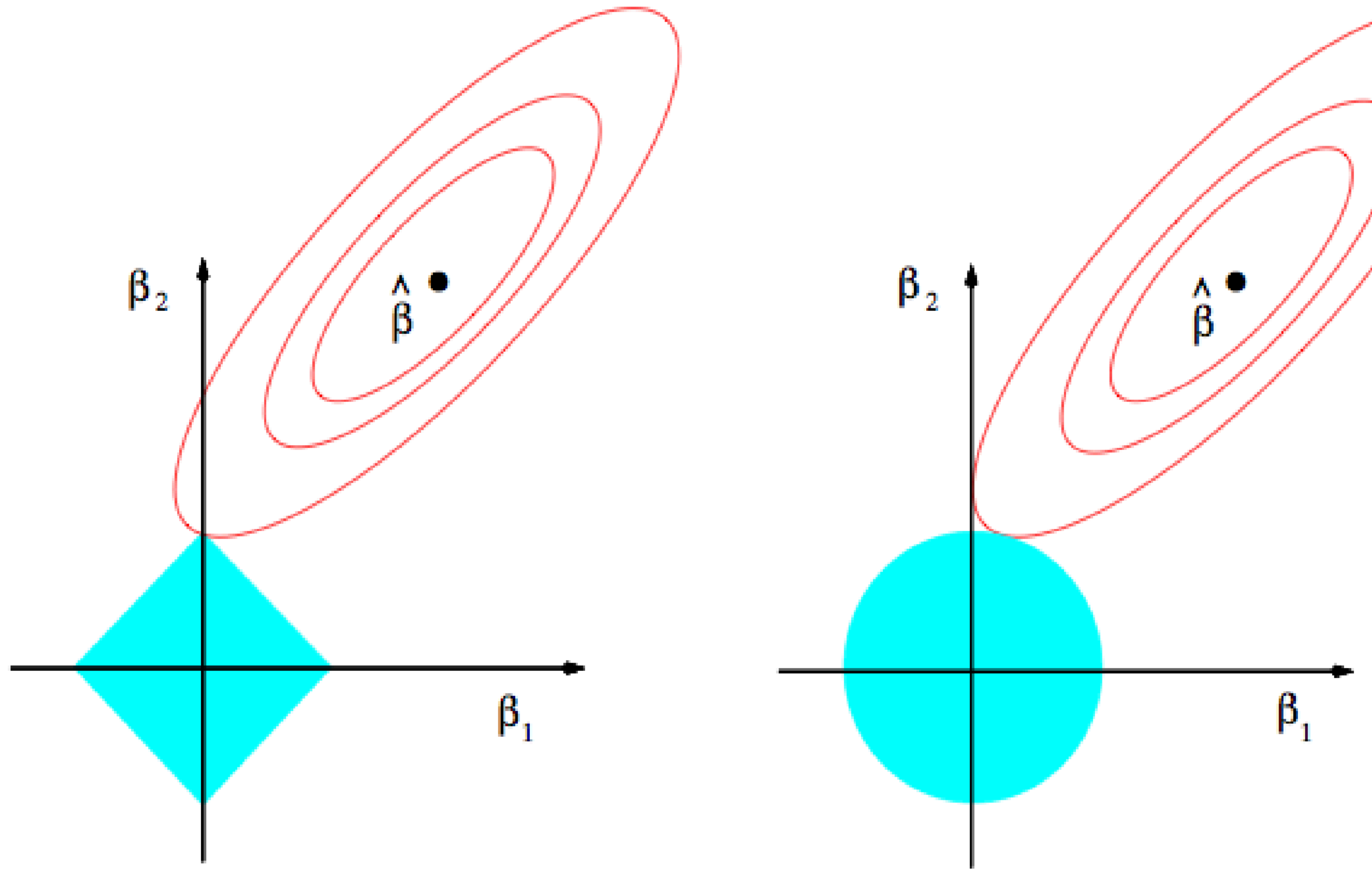

Figure 1: Comparing contour plots for LASSO (left) vs. ridge regression (right).

LASSO summary

- Relax the l_0 constraint to a l_1 constraint, which also encourages sparsity
- l_1 is not differentiable everywhere, but we can still find a closed solution using coordinate descent
- The LASSO can be applied to other problems other than linear regression! E.g., SVM.

Outline

1. Motivation for sparsity
2. The LASSO
3. **Matching pursuit**

Matching pursuit

- Strategy of LASSO: find another optimization problem that has similar properties but we can solve
- Strategy of MP: Keep the l_0 constraint and greedily improve solutions that satisfy it

Overview of matching pursuit

$$\min_{w} ||Xw - y||_2^2$$

$$\text{s.t. } ||w||_0 \leq k$$

1. Start with $w = \vec{0}$ (so it satisfies the constraint)
2. Iteratively update one component w_i at a time until the sparsity constraint $||w||_0 \leq k$ is violated
3. Pick which component to update by the one that minimizes the **residual** $r = ||y - Xw||_2^2$

Matching pursuit algorithm

Algorithm 2: Matching Pursuit

initialize the weights $\mathbf{w}^0 = \mathbf{0}$ and the residual $\mathbf{r}^0 = \mathbf{y} - \mathbf{X}\mathbf{w}^0 = \mathbf{y}$

while $\|\mathbf{w}\|_0 < k$ **do**

find the feature i for which the length of the projected residual onto $\mathbf{x}_i$ is maximized:

$$i = \arg\min_j \left(\min_\nu \|\mathbf{r}^{t-1} - \nu \mathbf{x}_j\|\right) = \arg\max_j \frac{|\langle \mathbf{r}^{t-1}, \mathbf{x}_j \rangle|}{\|\mathbf{x}_j\|}$$

update the i'th feature entry of the weight vector:

$$w_i^t = w_i^{t-1} + \frac{\langle \mathbf{r}^{t-1}, \mathbf{x}_i \rangle}{\|\mathbf{x}_i\|^2}$$

update the residual vector: $\mathbf{r}^t = \mathbf{y} - \mathbf{X}\mathbf{w}^t$