

# Lasso, Ridge, and Elastic Net

David Rosenberg

New York University

February 7, 2017

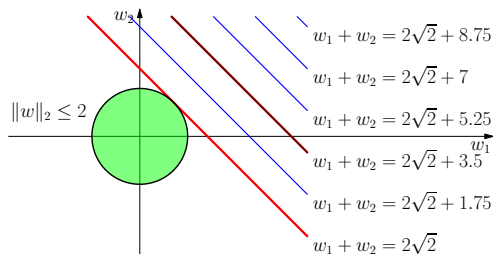
## Linearly Dependent Features

## Suppose We Have 2 Equal Features

- Input features:  $x_1, x_2 \in \mathbf{R}$ .
- Outcome:  $y \in \mathbf{R}$ .
- Linear prediction functions  $f(x) = w_1 x_1 + w_2 x_2$
- Suppose  $x_1 = x_2$ .
- Then all functions with  $w_1 + w_2 = k$  are the same.
  - give same predictions and have same empirical risk

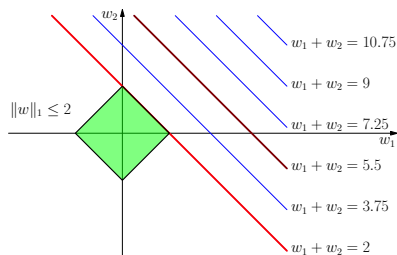
What function will we select if we do ERM with  $\ell_1$  or  $\ell_2$  constraint?

# Equal Features, $\ell_2$ Constraint



- Suppose the line  $w_1 + w_2 = 2\sqrt{2} + 3.5$  corresponds to the empirical risk minimizers.
- Empirical risk increase as we move away from these parameter settings
- Intersection of  $w_1 + w_2 = 2\sqrt{2}$  and the norm ball  $\|w\|_2 \leq 2$  is ridge solution.
- Note that  $w_1 = w_2$  at the solution

# Equal Features, $\ell_1$ Constraint

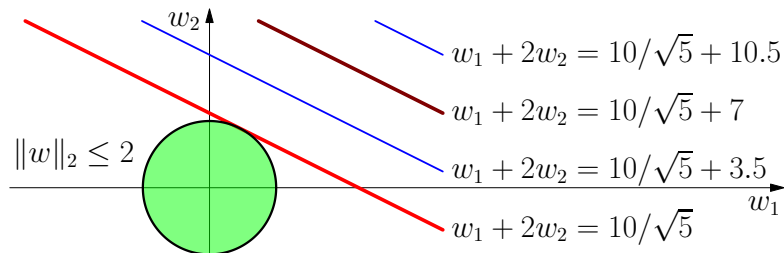


- Suppose the line  $w_1 + w_2 = 5.5$  corresponds to the empirical risk minimizers.
- Intersection of  $w_1 + w_2 = 2$  and the norm ball  $\|w\|_1 \leq 2$  is lasso solution.
- Note that the solution set is  $\{(w_1, w_2) : w_1 + w_2 = 2, w_1, w_2 \geq 0\}$ .

# Linearly Related Features

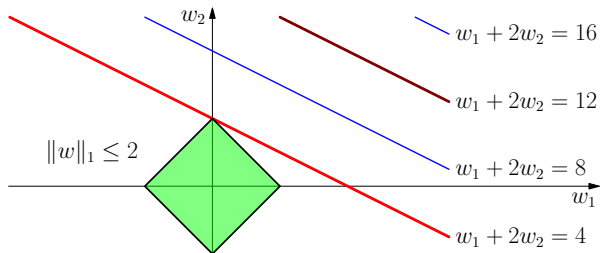
- Same setup, now suppose  $x_2 = 2x_1$ .
- Then all functions with  $w_1 + 2w_2 = k$  are the same.
- give same predictions and have same empirical risk  
What function will we select if we do ERM with  $\ell_1$  or  $\ell_2$  constraint?

# Linearly Related Features, $\ell_2$ Constraint



- $w_1 + 2w_2 = 10/\sqrt{5} + 7$  corresponds to the empirical risk minimizers.
- Intersection of  $w_1 + 2w_2 = 10\sqrt{5}$  and the norm ball  $\|w\|_2 \leq 2$  is ridge solution.
- At solution,  $w_2 = 2w_1$ .

# Linearly Related Features, $\ell_1$ Constraint



- Intersection of  $w_1 + 2w_2 = 4$  and the norm ball  $\|w\|_1 \leq 2$  is lasso solution.
- Solution is now a corner of the  $\ell_1$  ball, corresponding to a sparse solution.

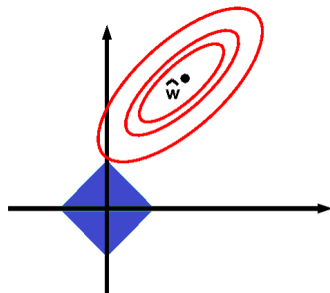


# Linearly Dependent Features: Take Away

- For identical features
  - $\ell_1$  regularization spreads weight arbitrarily (all weights same sign)
  - $\ell_2$  regularization spreads weight evenly
- Linearly related features
  - $\ell_1$  regularization chooses variable with larger scale, 0 weight to others
  - $\ell_2$  prefers variables with larger scale – spreads weight inversely proportional to scale

# Empirical Risk for Square Loss and Linear Predictors

- Recall **our discussion** of linear predictors  $f(x) = w^T x$  and square loss.
- Sets of  $w$  giving same empirical risk (i.e. level sets) formed ellipsoids around the ERM.



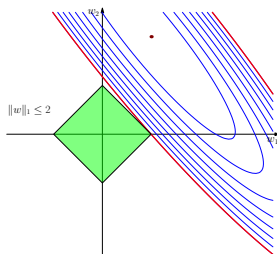
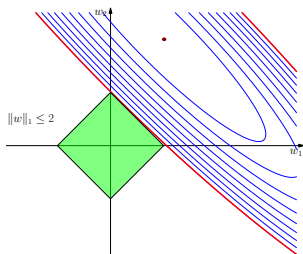
- With  $x_1$  and  $x_2$  linearly related, we get a degenerate ellipse.
- That's why level sets were lines (actually pairs of lines, one on each side of ERM).

KPM Fig. 13.3

## Correlated Features – Same Scale

- Suppose  $x_1$  and  $x_2$  are highly correlated and the same scale.
- This is quite typical in real data, after normalizing data.
- Nothing degenerate here, so level sets are ellipsoids.
- But, the higher the correlation, the closer to degenerate we get.
- That is, ellipsoids keep stretching out, getting closer to two parallel lines.

# Correlated Features, $\ell_1$ Regularization



- Intersection could be anywhere on the top right edge.
- Minor perturbations (in data) can drastically change intersection point – very unstable solution.
- Makes division of weight among highly correlated features (of same scale) seem arbitrary.
  - If  $x_1 \approx 2x_2$ , ellipse changes orientation and we probably hit a corner.

# A Very Simple Model

- Suppose we have one feature  $x_1 \in \mathbf{R}$ .
- Response variable  $y \in \mathbf{R}$ .
- Got some data and ran least squares linear regression.
- The ERM is

$$\hat{f}(x_1) = 4x_1.$$

- What happens if we get a new feature  $x_2$ ,
  - but we always have  $x_2 = x_1$ ?

# Duplicate Features

- New feature  $x_2$  gives no new information.
- ERM is still

$$\hat{f}(x_1, x_2) = 4x_1.$$

- Now there are some more ERM:

$$\hat{f}(x_1, x_2) = 2x_1 + 2x_2$$

$$\hat{f}(x_1, x_2) = x_1 + 3x_2$$

$$\hat{f}(x_1, x_2) = 4x_2$$

- What if we introduce  $\ell_1$  or  $\ell_2$  regularization?

# Duplicate Features: $\ell_1$ and $\ell_2$ norms

- $\hat{f}(x_1, x_2) = w_1 x_1 + w_2 x_2$  is an ERM iff  $w_1 + w_2 = 4$ .
- Consider the  $\ell_1$  and  $\ell_2$  norms of various solutions:

$w_1$	$w_2$	$\ w\ _1$	$\ w\ _2^2$
4	0	4	16
2	2	4	8
1	3	4	10
-1	5	6	26

- $\|w\|_1$  doesn't discriminate, as long as all have same sign
- $\|w\|_2^2$  minimized when weight is spread equally
- Picture proof: Level sets of loss are lines of the form  $w_1 + w_2 = c \dots$

## Correlated Features and the Grouping Issue



## Example with highly correlated features

- Model in words:
  - $y$  is a linear combination of  $z_1$  and  $z_2$
  - But we don't observe  $z_1$  and  $z_2$  directly.
  - We get 3 noisy observations of  $z_1$ .
  - We get 3 noisy observations of  $z_2$ .
- We want to predict  $y$  from our noisy observations.

---

Example from Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

## Example with highly correlated features

- Suppose  $(x, y)$  generated as follows:

$$\begin{aligned} z_1, z_2 &\sim \mathcal{N}(0, 1) \text{ (independent)} \\ \varepsilon_0, \varepsilon_1, \dots, \varepsilon_6 &\sim \mathcal{N}(0, 1) \text{ (independent)} \\ y &= 3z_1 - 1.5z_2 + 2\varepsilon_0 \\ x_j &= \begin{cases} z_1 + \varepsilon_j/5 & \text{for } j = 1, 2, 3 \\ z_2 + \varepsilon_j/5 & \text{for } j = 4, 5, 6 \end{cases} \end{aligned}$$

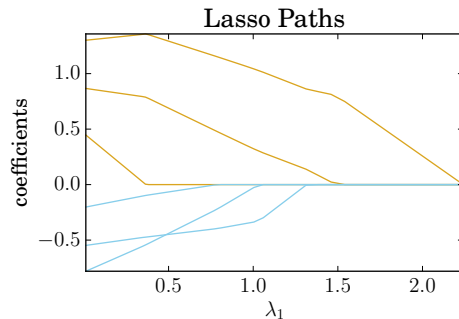
- Generated a sample of  $(x, y)$  pairs of size 100.
- Correlations within the groups of  $x$ 's were around 0.97.

---

Example from Section 4.2 in Hastie et al's *Statistical Learning with Sparsity*.

## Example with highly correlated features

- Lasso regularization paths:



- Lines with the same color correspond to features with essentially the same information
- Distribution of weight among them seems almost arbitrary

# Hedge Bets When Variables Highly Correlated

- When variables are highly correlated (and same scale, after normalization),
  - we want to give them roughly the same weight.
- Why?
  - Let their error cancel out
- How can we get the weight spread more evenly?

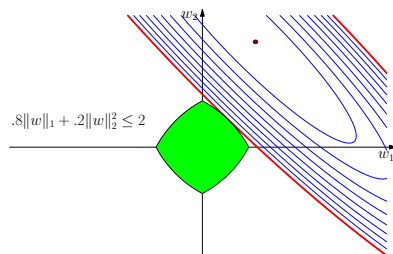
# Elastic Net

- The **elastic net** combines lasso and ridge penalties:

$$\hat{w} = \arg \min_{w \in \mathbf{R}^d} \frac{1}{n} \sum_{i=1}^n \{w^T x_i - y_i\}^2 + \lambda_1 \|w\|_1 + \lambda_2 \|w\|_2^2$$

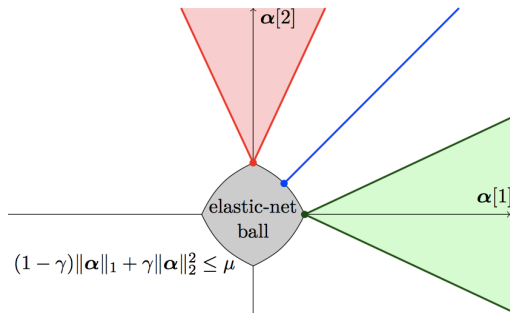
- We expect correlated random variables to have similar coefficients.

# Highly Correlated Features, Elastic Net Constraint



- Elastic net solution is closer to  $w_2 = w_1$  line, despite high correlation.

# Elastic Net - “Sparse Regions”



- Suppose design matrix  $X$  is orthogonal, so  $X^T X = I$ , and contours are circles (and features uncorrelated)
- Then OLS solution in green or red regions implies elastic-net constrained solution will be at corner

Fig from Mairal et al.'s *Sparse Modeling for Image and Vision Processing* Fig 1.9

# Elastic Net – A Theorem for Correlated Variables

## Theorem

<sup>a</sup>Let  $\rho_{ij} = \widehat{\text{corr}}(x_i, x_j)$ . Suppose  $\hat{w}_i$  and  $\hat{w}_j$  are selected by elastic net, with  $y$  centered and predictors  $x_1, \dots, x_d$  standardized. If  $\hat{w}_i \hat{w}_j > 0$ , then

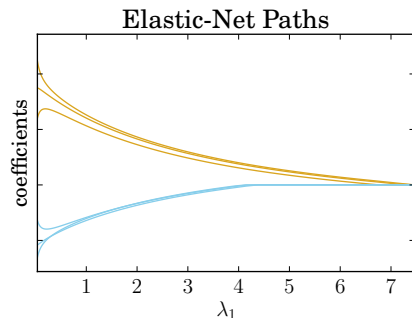
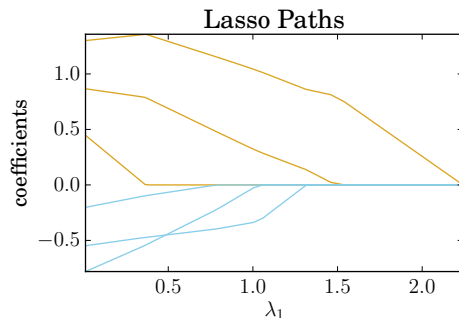
$$|\hat{w}_i - \hat{w}_j| \leq \frac{\|y\| \sqrt{2}}{\lambda_2} \sqrt{1 - \rho_{ij}}.$$

---

<sup>a</sup>Theorem 1 in “Regularization and variable selection via the elastic net”:  
[https://web.stanford.edu/~hastie/Papers/B67.2%20\(2005\)%20301-320%20Zou%20&%20Hastie.pdf](https://web.stanford.edu/~hastie/Papers/B67.2%20(2005)%20301-320%20Zou%20&%20Hastie.pdf)



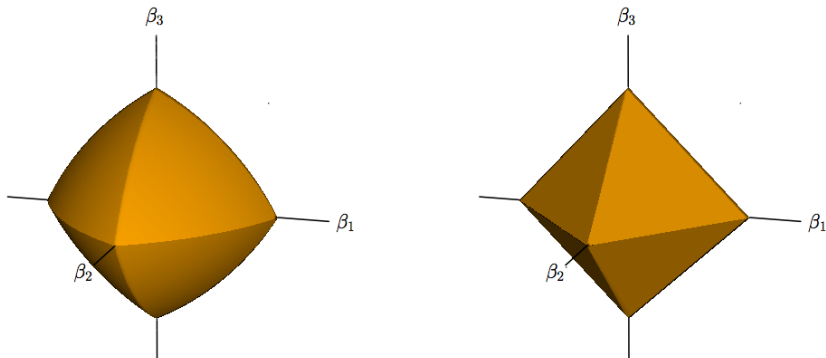
# Elastic Net Results on Model



- Lasso on left; Elastic net on right.
- Ratio of  $\ell_2$  to  $\ell_1$  regularization roughly 2 : 1.

## Extra Pictures

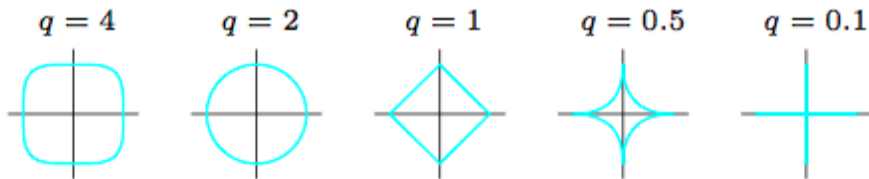
# Elastic Net vs Lasso Norm Ball



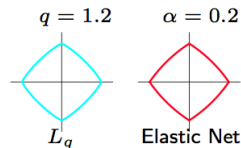
From Figure 4.2 of Hastie et al's *Statistical Learning with Sparsity*.

# The $(\ell_q)^q$ Constraint

- Generalize to  $\ell_q$  :  $(\|w\|_q)^q = |w_1|^q + |w_2|^q$ .
- Note:  $\|w\|_q$  is a norm if  $q \geq 1$ , but not for  $q \in (0, 1)$
- $\mathcal{F} = \{f(x) = w_1 x_1 + w_2 x_2\}$ .
- Contours of  $\|w\|_q^q = |w_1|^q + |w_2|^q$ :



# $\ell_{1.2}$ vs Elastic Net



**FIGURE 3.13.** *Contours of constant value of  $\sum_j |\beta_j|^q$  for  $q = 1.2$  (left plot), and the elastic-net penalty  $\sum_j (\alpha \beta_j^2 + (1 - \alpha) |\beta_j|)$  for  $\alpha = 0.2$  (right plot). Although visually very similar, the elastic-net has sharp (non-differentiable) corners, while the  $q = 1.2$  penalty does not.*