The Lasso Regression

Stands for **least absolute shrinkage and selection operator** ,Same as Ridge Regression which penalize linear regression, but the main disadvantage of the ridge regression is it will shrink the coefficients but not set any of them to zero which can be a challenge, Since the resulting model contains all the predictors, So when **inference and interpretation** is needed **Lasso Regression** is desired

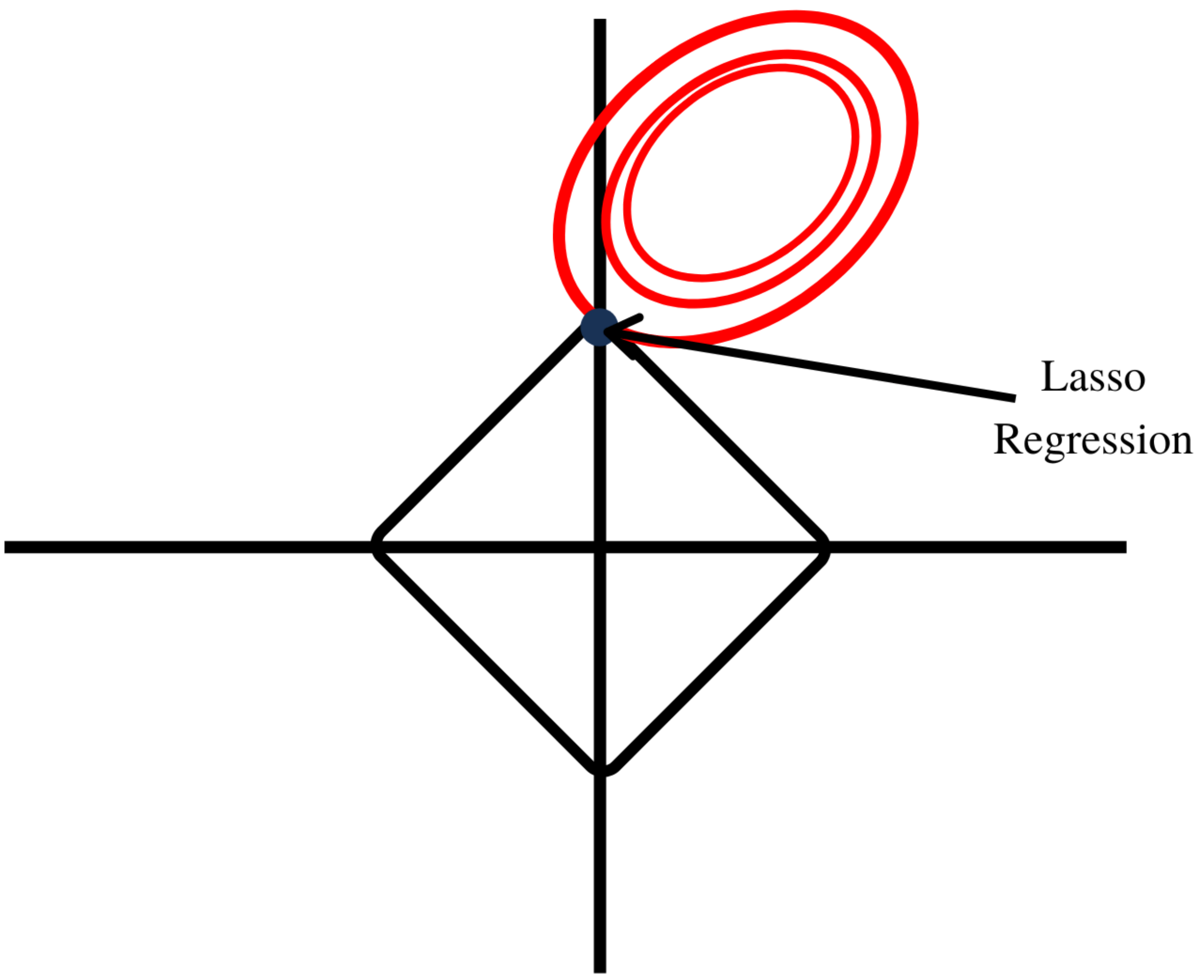
The **Ridge Regression** main motivation behind was to deal with :

- High Multicollinearity
- High Dimensionality
- Prediction Accuracy

And it used the **Squared Euclidean Norm** which is the L_2 Norm, they used it for an arbitrary reason behind which lead for a consideration in other Norms such as L_1 which is called **The Lasso Regression**

Lasso Vs Ridge

	Lasso	**Ridge**
Norm	Uses the L_1 Norm	Use the L_2 Norm
Penalty Term	$\lambda \sum \|\beta_j\|$	$\lambda \sum \beta_j^2$
Effect	Can set coefficients all the way to zero	Shrinks coefficients towards zero, never set them to **zero**
Use Case	Better performance and interpretability, and feature selection	Accurate predictions, prevent overfitting
Geometry	Circle or a hypersphere	diamond shape, often solution lies at a corner



Lasso Regression

It's introduce a penalty term same as the Ridge Regression but in the L_1 Norm which uses :

$$f_{pen}(\beta, \lambda) = \lambda_1 \|\beta\|_1$$

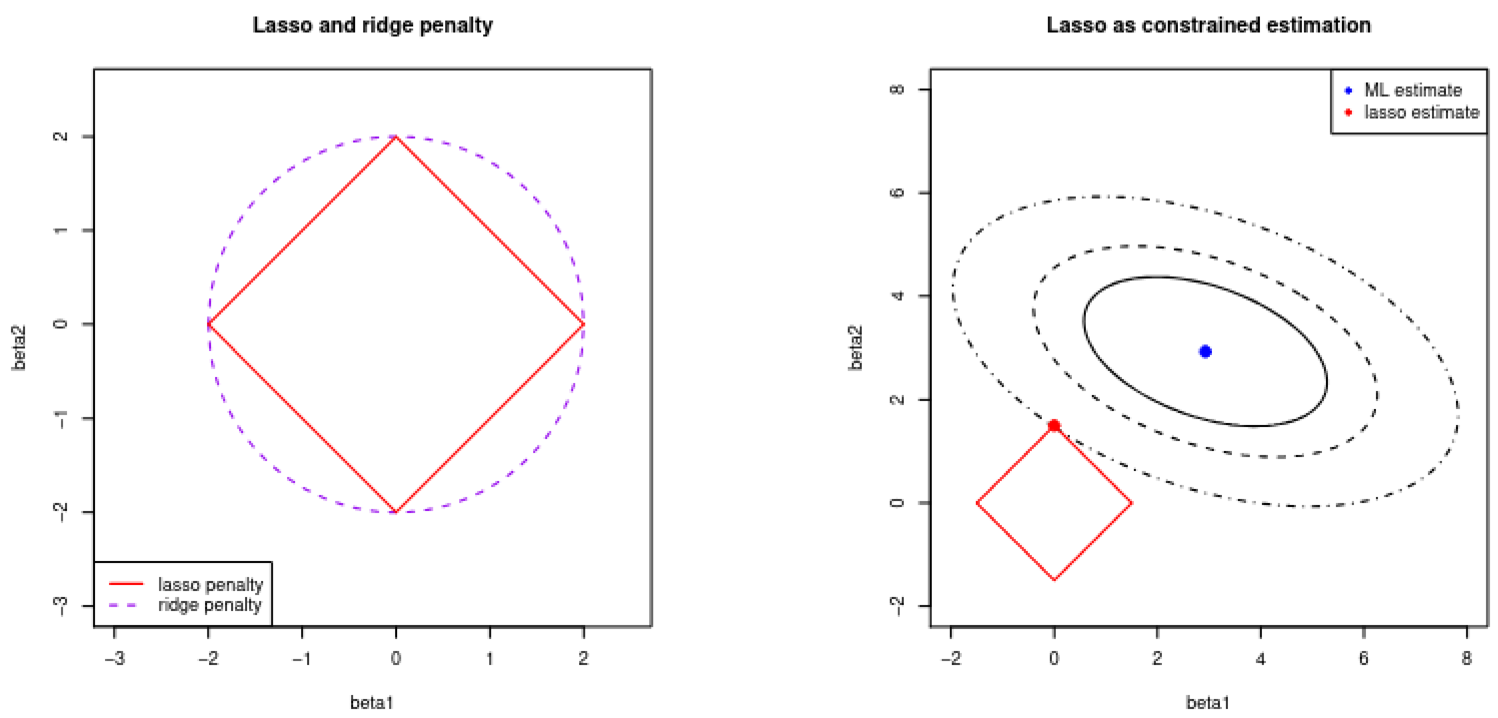
Which give us the **Lasso Cost Function**

$$\mathcal{L}_{\text{lasso}}(\beta;\lambda) = \|Y - X\beta\|_2^2 + \lambda_1\|\beta\|_1 = \sum_{i_1}^{n}(Y_i - X_i\beta)^2 + \lambda_1\sum_{j=1}^{p}|\beta_j|$$

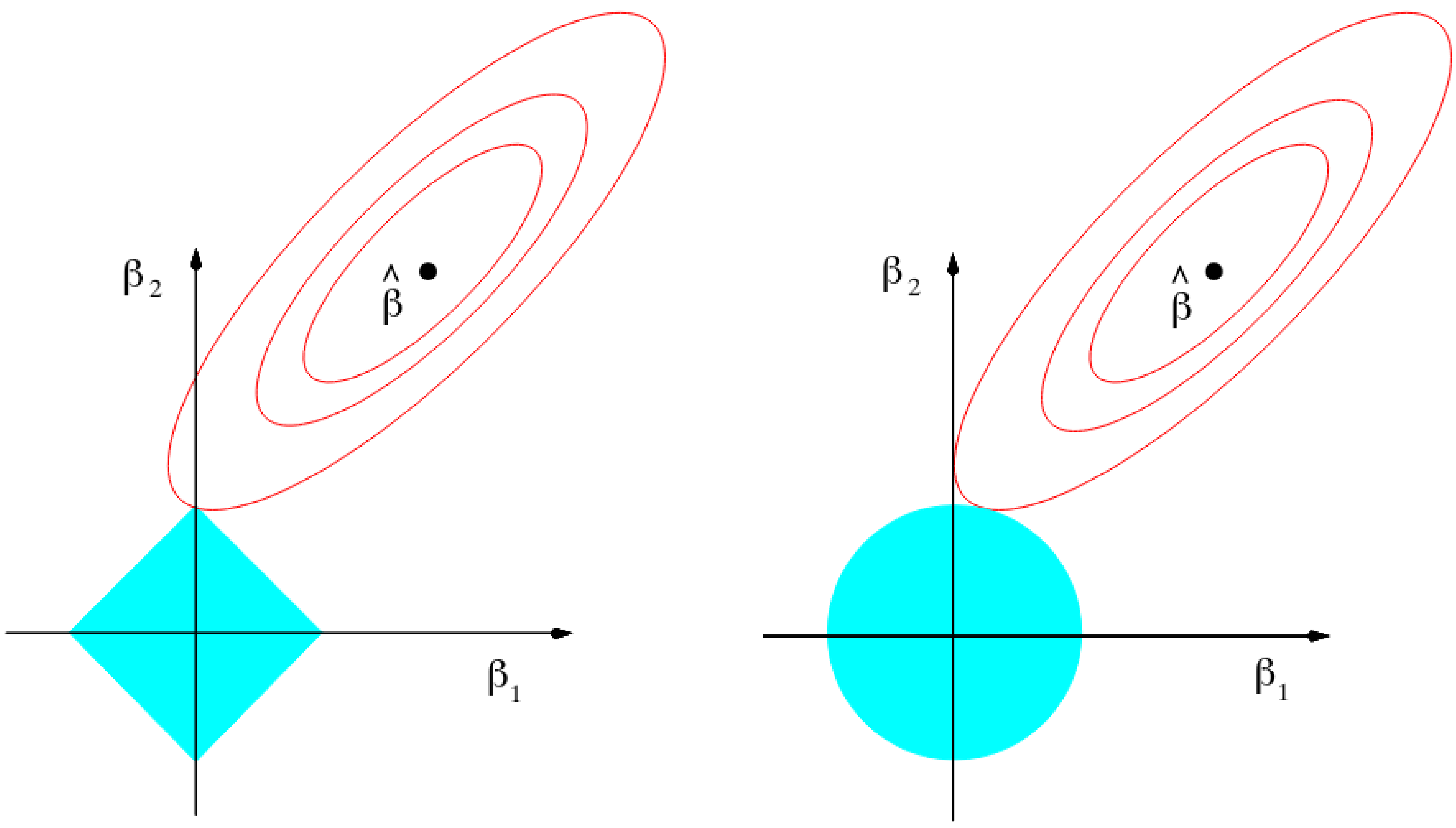
- Contains the **Least Squares** and **Regularization Term**
- The **Least Squares** term is not strictly convex due to high dimensionality
- The **absolute value** function is convex
- Which means the lasso loss function is convex but not strict
- Absolute value doesn't have a solution at 0 so no close-form solution exist unlike **Ridge Regression**

Intuition Behind Lasso Regression

- The **Lasso** Shrinks the coefficients towards zero same as ridge regression
- The L_1 penalty forces some coefficients estimates $\hat{\beta}$ to be exactly zero
- The **Lasso Regression** results in a spare model which means a model that only involve subset of the variables



- The constraints of the **Lasso** falls on it's corners on the axes where on of the coefficients is equal to zero



Why Lasso Set Coefficients to Zero

The thing that explains why the Lasso set some coefficients to zero is the **KKT** subgradient conditions also know as **stationarity**

The **stationarity** condition states for a given dual variable pair the point x minimize the lagrangian $\mathcal{L}$, and for convex function it can be written as (more details about the Lagrangian and optimization in Convex Optimization):

$$0 \in \partial f(x) + \sum \lambda \partial g_i(x) + \sum v_i \partial h_i(x)$$

Given the lasso problem :

$$\min_\beta \frac{1}{2n} \|Y - X\beta\|_2^2 + \lambda \|\beta\|_1$$

With $\lambda > 0$

Applying the **KKT stationarity** condition

$$0 \in \frac{1}{n} X^\intercal (Y - X\hat{\beta}) + \lambda \partial \|\hat{\beta}\|_1$$

$$\frac{1}{n} X^\intercal (Y - X\hat{\beta}) + \lambda \partial \|\hat{\beta}\|_1 = 0$$

With

$$\partial \|\hat{\beta}\|_1 = \begin{cases} \text{sign}(\hat{\beta}_j) & ,\hat{\beta}_j \neq 0 \\ \in [-1, 1] & ,\hat{\beta}_j = 0 \end{cases}$$

$$\frac{1}{n} X^\intercal (Y - X\hat{\beta}) = -\lambda \partial \|\hat{\beta}\|_1 \equiv -\frac{1}{n} X^\intercal (Y - X\hat{\beta}) = -\lambda \partial \|\hat{\beta}\|_1$$

Sparsity

- if $\hat{\beta}_j \neq 0$, then $|\partial \|\hat{\beta}\|_1| = 1$

$$\frac{1}{n} X^\intercal (Y - X\hat{\beta}) = -\lambda \, \text{sign}(\hat{\beta}_j) \implies |\frac{1}{n} X^\intercal (Y - X\hat{\beta})| = \lambda$$

- if $\hat{\beta}_j = 0$, then $\partial \|\hat{\beta}\|_1 \in [-1, 1]$

$$|\frac{1}{n} X^\intercal (Y - X\hat{\beta})| = \lambda |\partial \|\hat{\beta}_j\|| \leq \lambda$$

  Therefore

$$|\frac{1}{n} X^\intercal (Y - X\hat{\beta})| \leq \lambda$$

  The **KKT** forced $\hat{\beta}_j = 0$ since it's smaller than λ, simply small residuals correlations are killed