Regularization (Shrinkage)

As the name Suggest this section is about **Regularizing** or **Constraints** the coefficients estimates $\hat{\theta}$ also noted as $\hat{\beta}$, These are the two methods discussed here :

- Ridge Regression L_1
- The Lasso L_2

Before diving into these methods, taking a look at the **Norms** will help understanding and intuition since they are derived from.

Norms

When thinking of geometric vectors intuitively the direction and length of the vector are first that comes to mind, Simply **Norm** is a function that assigns each vector x it's **length** $\|x\|$ or **magnitude**

- $\|\lambda x\| = |\lambda|\|x\|$
- $\|x + y\| \leq \|x\| + \|y\|$
- $\|x\| \geq 0$ and if $\|x\| = 0 \iff x = 0$

The L_p Norm

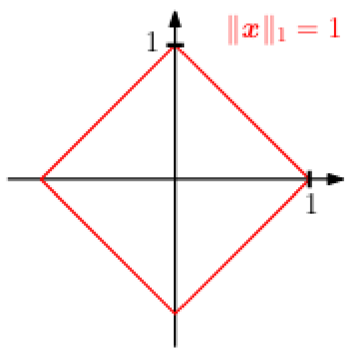
Also written as $\|x\|_p$, is defined as:

$$\|x\|_p = \sqrt[p]{\sum_{i=1}^{n} |x_i|^p}$$

with : $p > 0$ and x_i the **components** of x

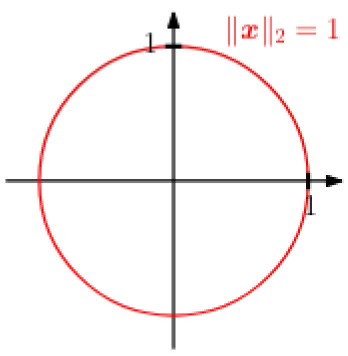
The L_1 Norm (Manhattan Norm)

$$\|x\|_1 = \sum_{i=1}^{n} |x_i|$$



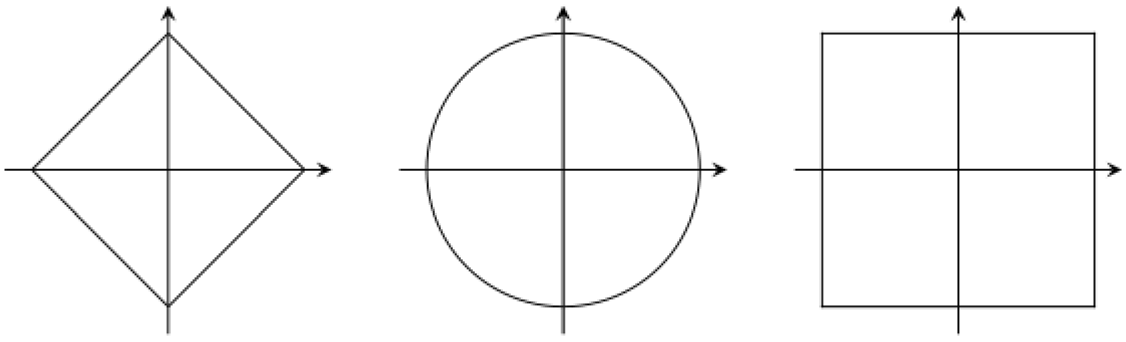
The L_2 Norm (Euclidean Norm)

$$\|x\|_2 = \sqrt{\sum_{i=1}^{n} x_i^2} = \sqrt{x^T x}$$



The L_∞ Norm

$$\|x\|_\infty = \max_i(|x_i|)$$



$p = 1$ $\quad$ $p = 2$ $\quad$ $p = \infty$

- Which results in a **square**

Ridge Regression

The Ridge Regression originally proposed to deal with the **Multicollinearity** in the predictors, The Ordinary Least Squares results in a **Best Linear Unbiased Estimators** β, Since highly correlated variables may cause the model to become **unstable** (abnormal high variances in $\hat{\beta}$) and accompanied by large values of the **estimates**.

The **Ridge** Solution suggest that we introduce **bias** into the coefficients estimates which lowers the **variance** introduced by the **collinearity** following the Bias-Variance Trade-Off

There is many cases where the number of **predictors** p exceed the number of observations or samples n, the **Design matrix** X is called high-dimensional which using Multiple Linear Regression yields no unique solutions, Since the number of Unknown p is larger than the number of equations n, and often high-dimensional data can lead to **Multicollinearity**

Why Ridge Regression is Used ?

- High Multicollinearity
- High Dimensionality
- Prediction Accuracy

Ridge Regression Estimator

It's was proven in Ordinary Least Squares that's the estimated value of β is given by :

$$\hat{\beta} = (X^\intercal X)^{-1} X^\intercal Y$$

- This estimator is only defined if the **Gram Matrix** is invertible
- When the **Design matrix** is high dimensional it's impossible to yield unique solutions
- When the Predictors of the **Design matrix** are highly correlated results in **unstable large estimates**
- Often overfits the data and picks noise

There is two ways to solve this invertibility problem :

- Moore-Penrose inverse : It's provides an **Unbiased** best linear estimator but suffers from overfitting and poor prediction capabilities since it yield a sensitive model (Higher variance)
- Ridge Regression estimator : It's **Biased** and shrunken toward zero with low variance

The Ridge Regression Estimator simply replace $X^\intercal X$:

$$X^\intercal X + \lambda I_{pp}$$

With :

- $\lambda \in [0, \infty)$ considered as a tuning parameter or **penalty parameter**, which solves the singularity by adding a positive matrix λI_{pp}

Results in the ridge regression estimator (coefficient estimate) :

$$\hat{\beta}(\lambda) = (X^\intercal X + \lambda I_{pp})^{-1} X^\intercal Y$$

Each value of the tuning parameter results in a different ridge regression estimator and the set of these estimates are called **Solution Path** or **Regularization Path**