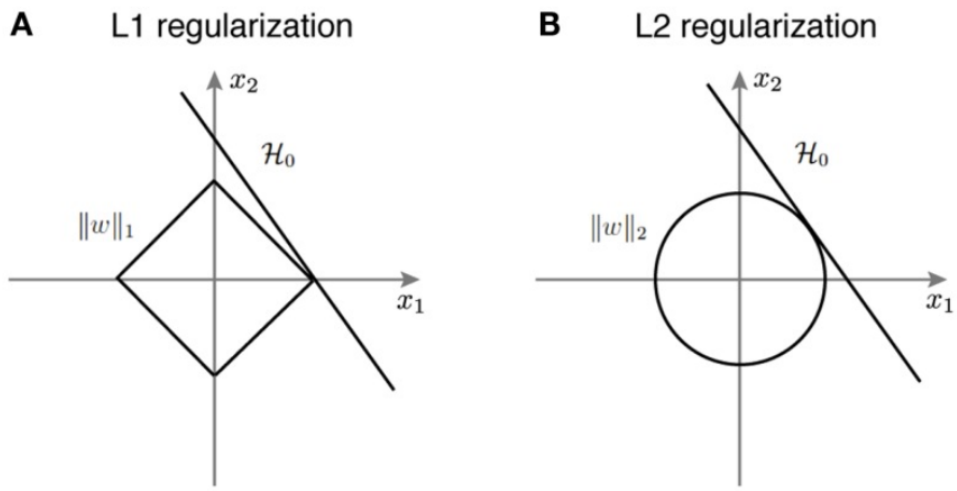
Regularization Implementation Issues

Regularization is a process of introducing additional information in order to solve an ill-posed problem or to prevent overfitting. In general, regularization is a technique that applies to objective function in ill-posed optimization problems. For instance, take a simple data set of two points. The simplest model is a straight line through the two points, or a first-degree polynomial. However, an infinite number of other models could also fit these points: second degree polynomials, third degree polynomials. . . and so on. Fitting a small amount of data will often lead to a complex, overfit model. A simpler model may be underfit and will perform poorly with predictions. Just because two data points fit a line perfectly doesn't mean that a third point will fall exactly on that line — in fact, it's highly unlikely. Simply put, regularization penalizes models that are more complex in favor of simpler models (ones with smaller regression coefficients), but not at the expense of reducing predictive power.

The following are the implementation issues associated with regularization:

- LASSO (least absolute shrinkage and selection operator) regularization or L1 regularization is a regression analysis method that performs variable selection and regularization. For cases with correlated features, LASSO can select only one feature from a group of correlated features. And, the selection is arbitrary in nature.
- LASSO regularization has a problem in variable selection problems where there are highly correlated features and there is a need to identify all the relevant ones.

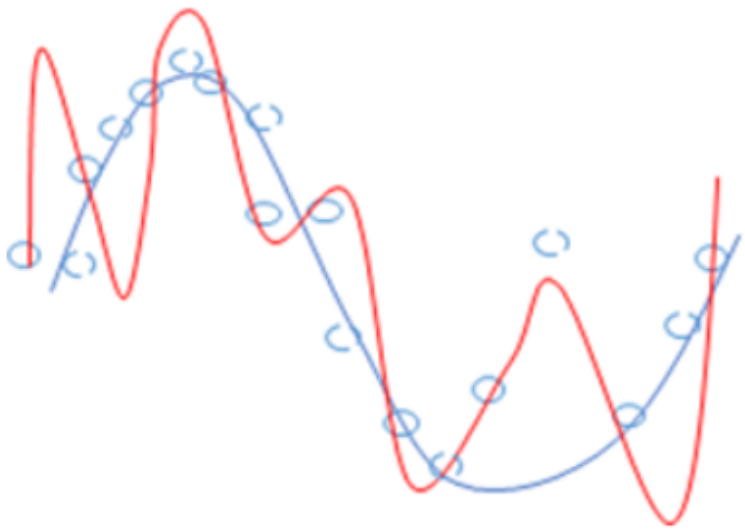
- In the case of a group of variables exhibiting high pairwise correlation, L1 does not care about which variable is selected.



The figure above depicts how different regularization terms lead to sparse and non-sparse solutions in a linear classifier. Image (A) illustrates how L1 regularization corresponds to the diamond shaped ball centered around the origin. Image (B) illustrates how L2 regularization corresponds to the spherical ball centered around the origin

- L1 regularization cannot aid in multicollinearity problems. It will just pick the feature with the largest correlation to the outcome.
- In small n (sample size) and large p (number of covariates) datasets, the LASSO method selects at most n variable before it saturates.
- If there are group variables present in the dataset (highly correlated between each other), LASSO tends to select one variable from each group while ignoring the others.
- L1 regularization uses information poorly when the relationship between predictive and target attributes is non-linear.
- The L1 norm is not differentiable and may require changes to learning algorithms in particular gradient-based learners.

- Ridge regularization also known as Tikhonov regularization or L2 regularization is the most commonly used method for regularization of ill-posed problems. It is unable to shrink coefficients to exactly zero and hence cannot perform variable selection.



In the figure above, the blue line is the true underlying model, the blue circles are noisy samples drawn from the model, and the red line is the regression model we learn from the training dataset. We can see that the learned model fits the training dataset perfectly, but it cannot generalize well on unseen data

- Ridge regression forces the learning coefficients to be lower but does not enforce them to be zero i.e. it will not get rid of the irrelevant features but rather minimize their impact on the trained model.
- L2 regularization, unlike subset selection, which will generally select models that involve just a subset of the variables, will include all the predictors in the final model.
- L2 regularization requires the number of instances to be larger than the number of attributes.

- L2 regularization does not have built-in feature selection like L1 regularization does. This is a result of the L1-norm, which tends to produce sparse coefficients. Suppose the model had 100 coefficients but only 10 of them are non-zero coefficients, this would mean that the remaining 90 predictors are not useful in predicting the target values. The L2 norm produces non-sparse coefficients and hence does not have this property.
- L2 regularization, being based on the minimization of a quadratic loss function, is sensitive to outliers (Maronna, 2011).
- L1 and L2 regularization require the variables to be standardized before using the regularization technique. Lasso regularization places constraints on the size of the coefficients associated with each variable. However, this value depends on the magnitude of each variable. The result of centering the variables means that there is no longer an intercept. This applies to ridge regression also.

Maronna, R. Robust Ridge Regression for High-Dimensional Data. 2011. Retrieved from: https://www.jstor.org